THE IMPACT OF DATA COLLECTION METHODS ON READING FLUENCY SCORES USING SECOND GRADE CURRICULUM-BASED MEASUREMENT READING (R-CBM) PROBES

A Dissertation
Presented to the
Faculty of the College of Education
Department of Educational Psychology
University of Houston

In Partial Fulfillment
of the Requirements for the Degree

Doctor of Philosophy

by

Dana E. Kelly

December, 2010

The impact of data collection methods on Reading Fluency scores using second-grade Curriculum-Based Measurement-Reading (R-CBM) probes.

A Dissertation for the Degree

Doctor of Philosophy

by

Dana E. Kelly

Approved by Dissertation Committee:

_____
Dr. Thomas Kubiszyn, Chairperson

_____
Dr. Weihua Fan, Committee Member

_____
Dr. Jacqueline Hawkins, Committee Member

_____
Dr. Milena Keller-Margulis, Committee Member

_____
Dr. Kimberly Schoger, Committee Member

_____
Dr. Robert K. Wimpelberg, Dean
College of Education
December, 2010

ACKNOWLEDGEMENT

To everyone who has helped make this possible,
(You know who you are)
I am sincerely grateful.

THE IMPACT OF DATA COLLECTION METHODS ON READING FLUENCY SCORES USING SECOND GRADE CURRICULUM-BASED MEASUREMENT READING (R-CBM) PROBES

An Abstract
of
A Dissertation Presented to the
Faculty of the College of Education
University of Houston

In Partial Fulfillment
of the Requirements for the Degree

Doctor of Philosophy

by

Dana E. Kelly

December, 2010

Kelly, Dana E. "The impact of data collection methods on Reading Fluency scores using second-grade Curriculum-Based Measurement-Reading (R-CBM) probes." Unpublished Doctor of Philosophy Dissertation, University of Houston, December, 2010.

Abstract

Due to changes in the Individuals with Disabilities Education Improvement Act (IDEIA, 2004), Curriculum-Based Measurement has expanded in its scope. This legislation established Response to Intervention (RtI) methods for use as prevention and early academic intervention to provide assistance to children who are having difficulty learning. According to this law, RtI data, such as Curriculum-Based Measurement for Reading (R-CBM) scores, can be used to determine which students are in need of more intensive interventions and may also be used in the diagnosis of specific learning disabilities, such as reading disabilities (IDEIA, 2004).

Currently, there are few evidence-based guidelines that inform R-CBM administration. For example, whether there are differences in R-CBM scores depending on the day of the week they are administered, and whether any such differences may be mitigated by the administration of three R-CBM probes as opposed to a single probe is unknown. Additionally, it is not known if there are significant differences in R-CBM scores if the median or the mean score (of three R-CBM probes) are utilized, and whether any such differences may be affected by the day of the week the probes are administered. The current study investigated the latter two issues.

The participants in the study were second-grade students who attended a local public school in south central United States. Data was collected for a period of six weeks during the spring semester. Essentially two questions were addressed. The first question addressed if a significant difference in reading fluency, as indicated by Words Read Correctly per Minute

(WRCM) exists depending on the number of probes given and the manner in which they are administered. Four administration conditions were examined: The first condition (Condition A) consisted of the median score of nine probes. For Condition A, three probes were administered to students three times each week. The other three conditions were contrived using the data from Condition A. Condition B consisted of the median score of three probes; each probe was collected on a different day of the week, such as Monday, Wednesday, and Friday. Condition C consisted of the median score of all three probes administered on the middle day of the week (Wednesday). Group D consisted of the first probe administered on Wednesdays, which simulated administering only one probe per week. To determine if there are differences in outcome depending upon the manner in which the probes are administered, a Two-way Repeated Measures ANOVA was used to analyze all of the weekly data.

Results suggested that there are differences in WRCM outcome depending upon the day of the week and the number of probes administered. The results indicated that overall, there is no difference in outcome if three probes are administered on one day or three probes spread out over three days during the week over the course of six weeks. Additionally, results showed that there was significantly greater R-CBM variability when one probe is administered in lieu of three probes or nine probes over the course of six weeks. Additional analyses using Repeated Measures ANOVAs were conducted to determine if there were differences between the conditions for each of the six weeks. Differences in outcome changed depending upon which week was examined, with the most consistency evident between Conditions A and B, Conditions A and C, and Conditions B and C. The most variability was seen between Conditions A and D and between Conditions B and D.

The second research question examined the difference in outcomes when the mean of scores is used as opposed to the median score. A Two-way Repeated Measures ANOVA was used, and results indicated that over time, there is no difference in outcome if the median or mean score is used across the four conditions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF GRAPHS

Chapter I

Introduction

Assessment that is used to adapt teaching to meet student needs is called formative assessment. A number of assessment approaches may be used to gain formative information about a student's progress, including curriculum-based assessment, general outcome measures, portfolio assessment, and informal teacher-developed tests. The frequent monitoring of students who are at-risk for academic failure or those who are currently below grade level can be assessed by using general outcome measures or GOMs. With GOMs, student performance on a common task is sampled over time to assess long-term growth and development. GOMs are deliberately intended *not* to be comprehensive. Instead, they measure direct, observable key skills that are representative of and related to important global outcomes, such as reading competence. Teachers can use data from GOMs to determine an individual student's progress and make modifications in instruction when necessary. School administrators can use aggregated GOM data to proactively improve the effectiveness of the instruction offered to all students. GOMs are highly sensitive to small, but important, changes in student performance. Because of these design features, GOMS can be administered frequently. Differences in scores are attributable to student growth, so educators can compare assessment results over time (Kaminski and Cummings, 2007).

Curriculum-Based Measurement or CBM is a type of GOM. According to Shinn (2002) CBM is a set of standardized and validated procedures comprised of short duration tests that are designed to measure academic progress in basic skills (Shinn, 2002). General education and special education teachers use CBM to help evaluate the effectiveness of their

instruction and interventions in the areas of reading, mathematics computation, spelling, and written expression. CBM measures are validated for use as formative evaluation as "dynamic indicators of basic skills," and are meant to function as "academic thermometers" to measure growth in areas relevant to school functioning (Shinn,1998, p. 1).

The use of CBM in schools has expanded as a result of changes in the Individuals with Disabilities Education Improvement Act (IDEIA, 2004). This legislation established Response to Intervention (RtI) methods for use as prevention and early academic intervention to provide assistance to children who are having difficulty learning. According to this law, RtI data, such as Curriculum-Based Measurement for Reading (R-CBM) scores may also be used in the diagnosis of specific learning disabilities, such as reading disabilities (IDEIA, 2004). Using R-CBM data to make eligibility decisions for the qualification of Special Education is what is referred to as a high-stakes decision. An example of a low-stakes decision would be determining whether a student receives additional instruction or intervention. R-CBM data is currently being used for both high-stakes and low-stakes decisions (Zirkel, 2010).

R-CBM involves the use of one-minute reading probes administered on a frequent basis; they are sensitive to small increments in learning (Shapiro, 2004). R-CBM probes are used in schools to measure student progress in basic reading skills (Shinn, 1998). According to Fuchs and Fuchs (2001), students are placed into three tiers, depending upon how much intervention and monitoring they need. Tier 1 should include an evidence-based instructional curriculum and universal screening (Fuchs & Fuchs, 2001). Within a three tier approach, CBM probes are used in school-wide tri-annual screenings (called Benchmarks) at the first tier—Tier 1. R-CBM data may be used to determine which students are progressing

typically and which students are in need of additional instruction. Students not making satisfactory progress are moved to Tier 2 where they are monitored more frequently and provided additional interventions. Performance on R-CBM measures may determine if students will receive additional interventions or not (Shapiro, 2004). Students who do not show evidence of progress after receiving research-based interventions are considered to be "non-responders," (Shapiro) and could be considered for special education services in Tier 3. Students who are categorized into Tier 3 would receive the most intensive intervention and monitoring of response to that intervention. In a three tier model, it is expected that approximately 80% of students would fall within Tier 1, approximately 15% in Tier 2, and 2%-7% in Tier 3 (Daly, Witt, Martens, & Dool, 1997).

Because IDEIA (2004) changed the manner in which learning disabilities are determined, R-CBM data can also inform special education eligibility decisions. Although there has been a recent increase in study of the psychometric properties of R-CBM (Marston, 1989; Merhens and Clarizio, 1993; Poncy, Skinner, and Axtell, 2005; Christ and Silberglitt, 2007), there remain a number of unanswered questions. One of these unanswered questions is whether there is a difference in outcomes when R-CBM probes are given on the same day or spread across days in the week. Another is whether a single probe yields different scores than the median of three probes, and another is whether the median and mean of three probes may differ.

Single CBM probes are not intended to be used for educational decision-making; rather performance on probes over time is analyzed. According to Christ and Ardoin (2009), measurement error is decreased as the number of probes given is increased. However, in terms of practical issues, it is not known if there is a difference in weekly Words Read

Correctly per Minute (WRCM) data if all of the probes are administered on the same day each week, or if they are administered on three different days each week. It is important to know how often and how many R-CBM probes need to be administered in order to accurately and reliably measure student progress. This is important, because decisions regarding the level of intervention (i.e., movement though the three tiers) for each student are made based on each student's data.

It is also not known if there is a difference in outcomes if the mean of the three weekly scores is used verses the median weekly score. The median score has historically been utilized, but comparisons of the median and means CBM score have not been located in the literature. It is important to know which measure best represents a student's performance. Because this CBM data can be used to make high-stakes decisions, it is important for the data collection process to be standardized to help decrease errors in measurement.

CHAPTER II

Review of the Literature

What is CBM

Curriculum-Based Measurement (CBM) is a set of standardized and validated procedures comprised of short duration tests that are designed to measure academic progress in basic skills (Shinn, 2002). General education and special education teachers use CBM to help evaluate the effectiveness of their instruction and interventions in the areas of reading, mathematics computation, spelling, and written expression (Shinn). CBM measures are validated for use as formative evaluation as "dynamic indicators of basic skills," (Shinn, 1998, p.5). When students are assessed at the end of an instructional program, teachers use a summative evaluation process. In contrast, formative evaluation involves continuous assessment during instruction; decisions are made based upon satisfactory or unsatisfactory progress (Shinn, 1998). The purpose of formative evaluation is to determine if the intervention was successful so that it can be modified or changed to increase the likelihood of positive outcomes. The key distinction between formative and summative evaluation is the role assessment plays in shaping the program for the student (Deno, 2002).

The primary purpose of CBM is formative assessment (Shinn, 1998). CBM can be used to help make decisions within a Problem-Solving Model, which promotes sequenced and differentiated assessment and decision making (Deno, 1989; Shinn, 1995). Within a problem-solving model, examination of an intervention could reveal that the intervention has not been successful in improving outcomes. At that point, an alternative approach would be used; the assessment continues throughout the process (Deno, 2002). There is research to support that the frequent use of CBM measures in a problem-solving model are valid and

effective predictors of academic performance and have concurrent validity with standardized, norm referenced measures. These procedures are primarily designed to assess students who are learning or who have difficulty learning basic skills (Shapiro, 2004).

According to Shinn (1998) the use of CBM procedures is dynamic in nature with regard to measurement, as it is sensitive to differences among individuals, as well as within individuals over a period of time. CBM can be viewed as an indicator of academic performance in a basic skill that can provide an indication of how an individual performs in a broader, related domain. CBM is considered primarily for use for acquisition of basic skills which are the framework for learning of all other skills; it is also for use with low-performing students who have not mastered the basic skills in the general education setting and those who are receiving services in special education. The primary group of students that would benefit from using CBM are those who are acquiring basic skills and those students "having specific achievement difficulties within specific curriculum skills areas" (Shinn, 1998, p. 16).

Curriculum-Based Measures in Reading (R-CBM) can be given frequently, take little time to administer (one minute for each probe), are sensitive to reading growth, and are well correlated (more than .60 in most studies) with reading comprehension tests (Deno, Mirkin, & Chiang, 1982; Hamilton & Shinn, 2003; Hintze, Callahan, Matthews, Williams, & Tobin, 2002; Shapiro, Edwards, Lutz, & Keller, 2004). R-CBM uses the number of words read correctly per minute (WRCM) to paint a picture of a student's overall reading proficiency (Deno, 1985). Because reading aloud is such a complex endeavor requiring coordination among several cognitive processes, it serves as an index of the student's general reading achievement and is extremely useful for monitoring a student's response to instruction (Fuchs, Fuchs, Hosp, & Jenkins, 2001).

History of CBM

Standardized, norm-referenced tests have been used for many years in the school setting to measure student performance, however, proponents of CBM say that norm-referenced assessments do not adequately inform educational programming for students. There is a great deal of debate regarding the use of standardized tests, and the following is a brief summary of the salient arguments against using standardized tests in schools. First, researchers argued that standardized tests lack content validity, because there may not be a direct match between the tests and the curriculum that is being taught. If content validity is low, the tests fail to provide a true measure of what the students have learned (Fuchs, Fuchs, & Maxwell, 1988). Next, standardized tests typically are not sensitive to minor changes in student progress, therefore they would be likely to show no change when there are only the small gains that are characteristically associated with day-to-day instruction (Marston, 1989). Third, standardized tests cannot be used effectively to monitor progress in day-to-day instruction, because they cannot be used in frequent intervals; they can be used to evaluate student outcomes only after the intervention has been completed, such as in a pre-test/post-test design (Marston, 1989).

Procedures for R-CBM were developed in the late 1970's and early 1980's in the context of a problem-solving model (Deno, 1985). Problem solving is viewed as an experiment for each student in which data is gathered and interpreted, and interventions are designed such that data can be produced for the evaluation of the intervention. There are generally two types of interventions in the schools: instruction that is delivered to all students, and special interventions that are designed for use with students not developing typically. Within this model, it is important that interventions are evaluated formatively so

that interventions can be changed or modified to increase the likelihood of achieving a goal (Deno, 2002).

Within the problem-solving model, each case is viewed as an experiment, which originally lent itself to a pre-test/post-test scenario. Students can be given pre-tests in order to determine their level of performance prior to an intervention. Upon completion of the intervention, students can be given a post-test to determine the amount of growth. Although this design is valuable, there are still questions that are unanswered, such as the effectiveness of the intervention. In addition, this design only confirms a discrepancy and is a static snapshot of the student's performance, and it does not illustrate the dynamic nature of learning (Marston, 1989).

According to Deno (2002), "Progress Monitoring" is a more effective method of assessing academic difficulties and evaluating the effectiveness of interventions. Progress monitoring refers to "direct and frequent observation of performance" of the skill(s) of the student. Students are measured repeatedly during baseline, intervention, and upon completion of the intervention, which produces a series of data points across time. These data points can be used to determine performance at certain intervals, but they can also be used to estimate trends in performance. The benefit of progress monitoring is that intervention effects can be closely monitored and changed if the intervention is not producing the desired effect (Deno, 2002).

Purpose of CBM

CBM allows teachers to measure baseline levels of performance on specific academic tasks and then to index proficiency and monitor progress in those areas. Teachers can use

CBM to help evaluate the effectiveness of interventions used for each individual student. CBM provides formative feedback to teachers so that alternative strategies and interventions can be used for students who do not adequately respond to initial instruction (Shapiro, 2004). Deficits in skill acquisition can be defined as the discrepancy between typical performance and atypical performance (Deno, 2002). The goal of CBM is to provide a framework to allow teachers to effectively help students gain proficiency in basic skills by systematically providing instruction, evaluating response to instruction, and changing instruction as needed (Shapiro, 2004). There is a body of literature that demonstrates the effectiveness of using CBM to improve student performance in basic skills, (Mirkin, Deno, Tindal, & Kuehnle, 1982; Fuchs, Deno, & Mirkin, 1984; Fuchs & Fuchs, 1986; Fuchs, Fuchs, & Hamlett, 1989) however there are many schools of thought in terms of how certain aspects of the data are collected and interpreted (Shapiro, 2004).

How CBM became part of the evaluation process

Discrepancy Model.

Prior to the Individuals with Disabilities Improvement Act of 2004 (IDEIA, 2004), Specific Learning Disabilities (SLD) were determined using a discrepancy model. In 1968, "specific learning disability" (LD) became a federally designated category of special education (U.S. Office of Education, 1968). That definition has remained substantively unchanged, and was reaffirmed in 1997 when Congressed reauthorized the Individuals with Disabilities Education Act [IDEA] (Public Law 105-17):

> "The term 'specific learning disability' means a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or

> written, which may manifest itself in imperfect ability to listen, think, speak, read, write, spell or do mathematical calculation. The term includes such conditions as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. Such term does not include a learning problem that is primarily the result of visual, hearing or motor disabilities, of mental retardation, of emotional disturbance, or of environmental, cultural, or economic disadvantage" (IDEA Amendments of 1997, PL105-17, 11 Stat. 37 [20 USC 1401(26)]).

The definition of an IQ-achievement discrepancy was introduced by Bateman (1965) as "an educationally significant discrepancy between estimated intellectual potential and actual level of performance related to basic disorders in the learning processes" (p. 220). This definition was not formally adopted by the federal government, however, the Bureau of Education for the Handicapped outlined procedures for LD identification that were related to this definition. The U.S. Office of Education (USOE; 1976) regulations read as follows:

> "A specific learning disability may be found if a child has a severe discrepancy between achievement and intellectual ability in one or more of several areas: oral expression, written expression, listening comprehension or reading comprehension, basic reading skills, mathematics calculation, mathematics reasoning, or spelling. A 'severe discrepancy' is defined to exist when achievement in one or more of the areas falls at or below 50% of the child's expected achievement level, when age and previous educational experiences are taken into consideration" (p. 52405).

In1977, the USOE stipulated regulations for the identification of SLD using a discrepancy model but disregarded the SLD formula. It was legislated that:

> "A team may determine that a child has a specific learning disability if: (1) The child does not achieve commensurate with his or her age and ability in one or more of the areas…when provided with learning experiences appropriate for the child's age and ability levels; and (2) The team finds that a child has a severe discrepancy between achievement and intellectual ability in one or more of the following areas: (i) oral expression, (ii) listening comprehension, (iii) written expression, (iv) basic reading skill, (v) reading comprehension, (vi) mathematics calculation, or (vii) mathematics reasoning" (USOE, 1977, p. 65083).

When Congress reauthorized IDEA in 2004, the procedures used to identify children with specific learning disabilities were changed. IDEA 2004 says schools "shall not be required to take into consideration whether a child has a severe discrepancy between achievement and intellectual ability in oral expression, listening comprehension, written expression, basic reading skill, reading comprehension, mathematical calculation, or mathematical reasoning" (Section 1414(b)). The USOE described several reasons why discrepancy models should be abandoned. The IQ-discrepancy criterion is potentially harmful to students, because it results in delaying intervention until the student's achievement is sufficiently low that the discrepancy is achieved. For many students, identification as having an SLD occurs at an age when the academic problems may be too difficult to remediate with the most intense remediation efforts (Torgesen, et al., 2001). In addition, the "wait to fail" model does not lead to "closing the achievement gap" for most students placed in special education, because students are not getting interventions until they have failed. At that point they are likely too far behind in the basic skills upon which later academic skill development is based to catch up (Donovon & Cross, 2002).

Cognitive Processing Model.

Under the IDEIA 2004, a cognitive processing model may also be used to determine if a student is eligible for special education services as a student with a Specific Learning Disability. The law states that each state "May permit the use of other alternative research-based procedures for determining whether a child has a specific learning disability as defined in §300.8 (c)(10)" (IDEA, 20 U.S.C. §1414 (b)(6)(A). In terms of SLD identification, there is some consensus among professionals that certain psychological processing difficulties are involved in SLD. Some examples of those limitations include working memory capacity, phonological processing deficits, and auditory perception (Flanagan, Ortiz, & Alfonso, 2007). Examining processing deficits helps to identify a disorder in basic psychological processes, which is a component of the federal definition of SLD. A comprehensive evaluation would consist of measurement of specific psychological processes and would include assessment of academic measures in order to establish links between the psychological process and academic area of concern (Flanagan, Ortiz, & Alfonso, 2007). Perhaps the most salient argument for this type of approach is being able to translate the data obtained from cognitive/academic assessment to specific strategies and interventions that can be used in conjunction with the student's strengths (Fiorello, Hale, & Snyder, 2006).

Several points were made by Ofeish (2006) regarding the Cognitive Processing model. There are advantages to using this model for the high-stakes decision of placement in special education. For example, using the cognitive processing model allows for the determination that there is a disorder in one or more of the basic psychological processes, which is essential to meet the legal definition of a Specific Learning Disability. This model assumes the use of traditional, standardized tests. Standardized tests are norm referenced and can provide

information such as grade and age equivalence, which can be used to determine the severity of the disability (Ofeish, 2006). Based on results from standardized, norm-referenced testing, psychologists can determine which area or areas of academic achievement are related to deficits in certain areas of cognitive processing. Use of intelligence tests help discriminate between students with learning disabilities and students with learning difficulties due to other contributory factors. An essential component of a learning disability is a failure to achieve in one or more areas at a level that is consistent with other abilities. Formal testing can be used to demonstrate whether or not the student possesses cognitive impairments in areas unrelated to the disability (Schrank et al., 2005).

There are other advantages to using standardized, norm-referenced tests. Holdnack & Weiss (2006) argue that the use of standardized tests assures that the disability is not situation specific (the student is low-functioning compared to children in one school, but in another school he/she would be considered average). These types of tests also allow for an in-depth understanding of a child's cognitive strengths and weaknesses (Holdnack & Weiss, 2006). Individual cognitive and neuropsychological assessments can inform instructional efforts that can then be systematically developed and evaluated through ongoing progress monitoring (Kavale, Kaufman, Naglieri, & Hale, 2005).

Although the cognitive processing model is supported, there are some researchers who suggest there are still some issues with the model. According to Dykeman (2006), this model is also a "wait to fail" model, where students only receive services once they are significantly behind. In order to serve students, there must be evidence of "educational need," which translates to poor performance for a given time period. This is an important point, because students may not receive the assistance they need unless and until they are identified as a

student in need of special education services. As stated earlier, it may be difficult for students to catch up if they have fallen too far behind (Donovon & Cross, 2002).

CBM has also been criticized for being a "wait to fail" model. A number of relevant issues were identified by Reynolds & Shaywitz (2009) as follows. In the case of using discrepancy models for identification of SLD, each Local Education Agency (LEA), in other words a local school, is free to devise and use their own method of SLD determination. This posed a problem of inconsistencies among LEAs, because there were many interpretations and incarnations of the process. There now exists a similar situation with the Response to Intervention (RtI) process due to a number of factors, such as a lack of guidance in assessing whether an RtI has occurred and inconsistencies in data collection and measurement models. Due to varied applications, different results will be obtained, depending upon the model that is used. Reynolds & Shaywitz (2009) also noted that an RtI model does not provide guidance in terms of instruction after a child has failed to respond. One major function of a comprehensive assessment, which includes cognitive and achievement testing, is providing a profile of a student's strengths and weaknesses that could lead to more effective remediation. Because all children who fail to respond may do so for different reasons, a more comprehensive evaluation can shed light on the root causes of learning difficulties, particularly with reading (Reynolds & Shaywitz, 2009). Once the underlying causes are identified, specific, research-based interventions can be used to target the student's deficits.

Response to Intervention

CBM has expanded in its use due to changes in SLD identification in the Individuals with Disability Improvement Act (IDEIA, 2004). This legislation established Response to

Intervention (RtI) procedures for use as prevention and early intervention programs. According to this law, RtI data may also be used in the diagnosis of specific learning disabilities (IDEIA, 2004). IDEA 2004 states, "when determining whether a child has a specific learning disability ... a local educational agency shall not be required to take into consideration whether a child has a severe discrepancy between achievement and intellectual ability" ... a school "may use a process that determines if the child responds to scientific, research-based intervention as part of the evaluation procedures ..." (Section 1414(b)(6)). The U. S. Department of Education "strongly recommends" that schools use a response to intervention model that

> "…uses a process based on systematic assessment of the student's response to high quality, research-based general education instruction…that incorporates response to a research-based intervention…Identification models that incorporate response to intervention represent a shift in special education toward the goals of better achievement and behavioral outcomes for students identified with SLD…"

Because of changes in the law, schools are compelled to provide intervention to students without them having to first be identified as having a learning disability. CBM procedures can be used in the RtI process, which is the degree to which students respond to research-based interventions. Students who do not respond positively could be potentially eligible for special education services (Shapiro, 2004). Because of this, the scope of CBM is broadened to include providing information to educators for use in high-stakes decisions, such as placement into special education services, which was not the original intent of CBM (Shinn, 1998).

The "promise" of CBM

Shinn (2002) indicates that there are several important features of General Outcome Measures, such as R-CBM. They measure important signs of general achievement, not every aspect of achievement. They are to be administered, scored and interpreted in a standard manner, since they are considered to be standardized tests. They are considered to be reliable and valid in terms of psychometric properties, and educators can feel confident in accurate measures of performance. CBM measures are sensitive to changes over short time periods. Both qualitative information and quantitative information can be gathered from observation of the target behavior(s). CBM probes are short in duration to ensure they will not take a great deal of time from academic instruction. Finally, Shinn (2002) states that CBM assessments are "linked to decision making for promoting positive achievement with general education students and for Problem-Solving decision making with at-risk students or those in remedial programs like Title I and special education" (p.8). Because it has been legislated that CBM is to be used as part of the process in identification of SLD, it is important to know if CBM measures are robust psychometrically. Although Shinn (2002) states that the procedures are administered in a "standard way," no empirical support for any particular manner of administration was able to be located.

Validity and Reliability

Reliability refers to the extent to which assessments are consistent. The values for reliability coefficients range from 0 to 1.0. A coefficient of 0 means no reliability and 1.0 means perfect reliability. Since all tests have some error, reliability coefficients never reach 1.0. Generally, if the reliability of a standardized test is above .80, it is said to have very good

reliability; if it is below .50, it would not be considered a very reliable test. *Validity* refers to the accuracy of an assessment—whether or not it measures what it is supposed to measure. Even if a test is reliable, it may not provide a valid measure. Since teachers, parents, and school districts make decisions about students based on assessments, the validity inferred from the assessments is essential. Also, if a test is valid, it must be reliable. The following section discusses first validity and then reliability evidence as it relates to R-CBM.

Validity. Marston (1989) examined the technical adequacy of CBM by reviewing articles that examined the validity and reliability of CBM passages. With regard to validity evidence, Marston described several studies which compare R-CBM Probes to other standardized measures. Marston (1989) discussed the Deno, Mirkin, and Chiang (1982) article which found high correlations between student performance on measures of reading fluency from a passage and norm-referenced, criterion tests of reading. When scores on standardized measures, such as the Stanford Diagnostic Reading Test (Karlsen, Madden, & Gardner, 1975), the Woodcock Reading Mastery Test (Woodcock, 1973), and the Reading Comprehension subtest of the Peabody Individual Achievement Test (Dunn & Markwardt, 1970) were compared to performance on reading passages, correlation coefficients ranged from .73 to .91. Twelve other studies reviewed by Marston (1989) found correlation coefficients ranging from .63 to .90 between oral reading rates and norm-referenced tests that measure global reading skills. Another study cited by Marston (1989) investigated oral reading fluency as it relates to reading comprehension (Fuchs, Fuchs, and Maxwell, 1988). They found that there was a high correlation (.89) between two reading comprehension subtests of the Stanford Achievement Test and oral reading fluency.

Marston (1989) also cites a study that investigated the construct validity of one-minute CBM oral reading probes (Deno, Marston, Shinn, and Tindal, 1983). Construct validity was measured using strategies to find evidence of discriminant validity, longitudinal studies of reading growth, and treatment validity. Discriminant validity in this study is defined as, "the degree to which the reading measure distinguished intact groups that differed in their reading skills." They were able to use the scores from the reading probes to differentiate students with learning disabilities and those without.

More recent validity evidence was presented by Shapiro, Keller, Lutz, Santoro, & Hintze (2006), who examined the relationship between CBM probes and statewide achievement tests and other standardized achievement tests for third-, fourth-, and fifth-grade students in Pennsylvania. Data was collected in two school districts in the state and consisted of 617 students from one district and 475 students for the reading portion of the study. In terms of the reading probes, they found that performance on the R-CBM measures were good predictors of performance on the statewide, year-end tests. Performance on R-CBM probes collected in the Fall, Winter, and Spring were compared to the Pennsylvania System of School Assessment (PSSA), the Metropolitan Achievement Test—Eighth Edition (MAT-8), the Stanford Achievement Test—Ninth Edition (SAT-9), and the Stanford Diagnostic Reading Test (SDRT). Strong correlations (up to .70) were found between the Winter and Spring administrations of R-CBM and the PSSA, which indicates a strong relationship between R-CBM and the state curriculum standards. In addition, the predictive power was found to be 80%-93% in terms of correctly identifying students below the cut-off criteria for R-CBM norms falling below the cut-off criteria for the PSSA. In terms of correctly identifying those students whose performance on CBM was above the cut-off criteria and

who also scored above the cut-off criteria for the PSSA, the predictive power was 48%-68%. With regard to the other standardized measures used in the study, correlations between R-CBM and MAT-8 and SAT-9 were in the .70s. It was noted that correlations between R-CBM and subtests which measure reading comprehension ranged from .65 to .74.

Prior to the Shapiro et al. (2006) study, a review was conducted of ten studies that compared CBM measures and outcomes to standardized state assessments in eight states. Powell-Smith (2004) summarized studies which examined the relationships between oral reading fluency and scores on high-stakes, statewide examinations. Correlations between performance on oral reading fluency and statewide achievement tests ranged from .44 to .79, with averages between the .60 to .75 range. These types of studies suggest a consistent relationship between scores from R-CBM reading probes and standardized assessment measures.

Reliability. Reliability is important to examine as CBM is expanded. There are implications if the reliability of CBM is not robust. As stated earlier, a test is not considered valid if it is not reliable; reliability coefficients of .80 and above are reflective of good test reliability. Conditions of testing can affect the reliability of R-CBM scores. Derr and Shapiro (1989) examined the variability among scores of curriculum based assessment (CBA), which were similar to R-CBM measures. In their study, variation in student performance was measured using Analysis of Variance (ANOVA). The effect of different measurement conditions was examined, and three conditions were studied: who administered the probes; the physical location of the assessment; and whether the subject was told he/she would be timed or not. Over a five-week period, 26 third- and fourth-grade, general education students were administered CBA reading probes. Results from this study indicate

that conditions of testing have an effect on the data for reading fluency. Derr and Shapiro (1989) found discrepancies in scores when the teacher administered the assessment versus the school psychologist; when the assessment occurred in an office outside of the classroom versus in the classroom; and when the examinees believed they were being timed or untimed. The findings of this study are important, because it illustrates some of the factors that can influence the scores obtained by students on R-CBM oral reading probes. Considerations regarding setting need to be addressed when examining student performance.

Standard Error of Measurement

The standard error of measurement (SEM) is the standard deviation of hypothetical error scores of a particular distribution of scores. Because error scores are a hypothetical construct, there is a formula which estimates the SEM based on obtained scores:

$$SEM=SD\sqrt{1-r}$$

In this formula, "r" is the reliability coefficient, and SD is the standard deviation of the test. The SEM provides information that allows educators to estimate the range of scores within which a student's true score falls (i.e., the score that reflects a student's true level of achievement, free of any measurement or random error). SEM is related to the reliability of the test; as the reliability coefficient decreases, the SEM will increase, all other factors being equal. The higher the SEM for a test, the less confident one can be that the test reliably measures the construct (Kubiszyn, 2010). The following section contains some research that has been conducted that addresses SEM with R-CBM.

In 1993, Merhens and Clarizio identified the Standard Error of Measurement as a potential deficit in the psychometric reliability of CBM. The authors indicate that the

Marston (1989) article, which examines reliability and validity of R-CBM, finds adequate reliability and validity for R-CBM. Merhens and Clarizio (1993), however state that the groups used in some of the studies are too homogeneous, which could lend itself to higher reliability. Additionally, standard errors of measurement, which are affected by group variability and error of difference scores are not addressed by Marston (1989). A number of additional criticisms of CBM were addressed by Merhens and Clarizio, such as limited focus on basic skills, low applicability to higher grades, poor standardization, difficulty of implementing with fidelity, as well as psychometric considerations. The authors recommend using CBM procedures as a supplement to existing procedures for the identification of students with disabilities rather than as a replacement.

Poncy, Skinner, and Axtell (2005) investigated the reliability and standard error of measurement of words read correctly using R-CBM probes with a sample of 37 third-grade students. They wanted to determine the percentage of variance in WCPM scores that was the result of student skills, the difficulty of the passage, and measurement error. They also investigated the Standard Error of Measurement (SEM) when there was variability in the reading probes. In their study, they found that 10% of variance in student performance was due to probe variability, 81% was due to student performance, and 9% was due to unaccounted for sources of error. They also found that they were able to lower the variance due to differences in probes from 9% to 2% by using a field-test procedure. The researchers additionally examined SEM with regard to the number of probes given. SEM decreased as the number of probes increased. When only one probe was administered, the coefficient of generalizability was .90 with a SEM of 12 WCPM, while the SEM decreased to 4 WCPM when the coefficient of generalizability increased to .99 when nine probes were administered.

Additionally, they found that if they used field-tested probes that were within +/- 5 WCPM of the average WCPM, the number of probes needed was reduced from 9 to 5 in order to achieve SEM of 4. Based on their findings, they state that using a single probe during the screening process would be sufficient, however, they recommend using at least three to five probes to help make programming decisions, such as providing additional interventions and informing eligibility decisions. It was also determined that field testing procedures were more effective than readability formulas for reducing the SEM. The importance of this study is that it indicates that it may be difficult to accurately detect small differences in student performance because of the error associated with R-CBM probes, and underscores the need to administer a sufficient number of probes before important decisions are made. Reducing variability in the passages and giving more probes can reduce error, which can increase the confidence with which important intervention and/or placement decisions are made (Poncy, Skinner, and Axtell, 2005).

Christ and Silberglitt (2007) conducted a study in order to derive estimates for the standard error of measurement (SEM) of reading CBM. The purpose of the study was to estimate the likely magnitude of SEM for R-CBM and identify a range of likely SEM values so that appropriate levels of confidence can be used when interpreting data. The researchers used archival data which was collected data in four-week intervals in the fall, winter, and spring each year across eight years. Approximately 8200 first- through fifth-grade students participated. During each point in the year, three successive probes were administered each week, and all analyses were conducted on the median of the three R-CBM probes collected at each time period. They found that the median estimate of SEM was 10 WRCM, with a range of 5 to 15 WRCM. It was inferred that as the conditions of testing become less ideal (such as

testing in a noisy classroom verses testing in a quiet environment) and more factors influence variability, the SEM is likely to be higher. In addition, the more variance in the group, the larger the SEM will be. The variance among groups appeared to be less in the lower grades and increased in the higher grades, as student performance varied more. In terms of reliability, it was noted that student performance on R-CBM probes is affected not only by instructional effects, but also by measurement conditions and the number of probes administered. It was concluded that R-CBM progress monitoring needs to be conducted under highly standardized assessment conditions and with a sufficient number of probes before important decisions are made. Due to the high levels of SEM, especially when fewer than six probes were administered, the authors suggested interpretation of performance be accompanied by confidence intervals.

Recent research related to probe equivalence and probe selection

Because CBM requires the use of multiple equivalent probes to measure progress over time, it is important for probes to be equivalent in terms of difficulty to ensure that progress is being measured adequately. There has been some recent research questioning the equivalence of alternate forms of reading passages. Christ and Ardoin (2009) noted that R-CBM passages were historically taken from curriculum texts. However, due to inconsistencies in passage difficulty within curriculum materials, standard practice today is to use commercially-made R-CBM passage sets. Christ and Ardoin note that readability formulas have been used to determine passage equivalence, however research regarding readability formulas demonstrated that readability formulas were not a good predictor of student reading fluency rates. In their 2009 study, Christ and Ardoin examined four different

methods for selecting reading passages: random selection of passages; selection based on readability results; selection based on mean levels of student performance from field testing; and use of passages based on measurement procedures using Euclidean Distance (ED is the square root of the sum of squared differences for student performances across passages).

Their study included 46 second-grade and 42 third-grade subjects. Results indicated that passages based on readability formulas were slightly better than randomly selected passages in terms of consistency of student performance. Better consistency was seen in the latter two conditions—passages chosen from field testing and measurement procedures. Results were similar across both grades. It was concluded that field testing and performance analysis would likely produce better alternate forms of R-CBM passages. Interestingly, the researchers found that the difference in the mean WRCM between the easiest and most difficult passage was 46 WRCM at each grade level. Because R-CBM probes are used for screening, benchmarking, progress monitoring, and the identification of specific learning disabilities, such variability could lead to errors in data that can have serious implications for students.

In another study designed to generate research-based recommendations regarding R-CBM probes, Ardoin and Christ (2008) investigated methods for selecting and using probes for universal screening. They wanted to examine the potential effect of using alternate probe sets on decisions made for educational programming. 86 second-grade students were assessed through a benchmark/screening process during the fall, winter, and spring using passages from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), which is a commercially-developed set of grade-level R-CBM passages. In this study, students were administered a single probe followed by three benchmark probes during the fall semester.

During the winter benchmark session, students were administered the same four probes that were used in the fall, in addition to a different set of benchmark probes. During the spring administration, students were again given the initial four probes, along with a different set of benchmark probes.

Ardoin and Christ (2008) found that reliability and validity coefficients were above .90 in most cases and the mean levels of WRCM were relatively stable across each administration. Alternate form reliabilities were high within each session. Coefficients were slightly higher when the mean score of three probes was used as compared to the single-probe administration (.97 vs. .94). Their study found support for the recommendation that using one probe is adequate for measuring reading levels for universal screening, however they caution the use of a single measurement due to potential variance, such as student interest. They state: "The magnitude of SEM in a typical population is likely to approximate 7 WRCM (range = 5 to 7) when the reliability of measurement is .97. It is likely to approximate 10 WRCM (range = 7 to 10) when the reliability of measurement is .94" (p. 120-121). In terms of benchmarking data, use of the same three probes for fall, winter, and spring rendered the most reliable scores when the median score was used (Ardoin & Christ, 2008). This finding is particularly important, because it highlights the need for additional exploration of other measures which could, potentially be more reliable (e.g. mean WRCM scores). Additionally, Ardoin and Christ's study found that depending on which probe or set of probes were used, scores obtained yielded different percentages of students who were identified as being discrepant in level of reading and rate of growth (2008). This study is an example of how difference in data collection methods produced different outcomes, which is an area that needs more empirical support.

Another important finding of this study is that median estimates were more robust in terms of estimating growth than single-probe administration (Ardoin & Christ, 2008). The difference in outcomes for this portion of the study is germane to the present topic, because more investigation is needed in the area of probe administration as it relates to student performance. In addition, Ardoin and Christ (2008) found that when addressing student growth, difference scores from two universal screenings is not sufficient, particularly for students found to be discrepant in level and rate. They also found that growth rates may not be the same from semester to semester. The researchers suggest that students receiving progress monitoring should not be compared to annual growth rates, because there is no research that compares progress monitoring data to benchmark data (Ardoin and Christ 2008). These findings are important, because there had been no prior research to guide this process.

Francis, Santi, Barr, Fletcher, Varisco, and Foorman conducted a study that examined the effects of passage and presentation order of R-CBM passages during progress monitoring of 134 second-grade students (2008). DIBELS oral reading fluency (Good & Kaminski, 2002a) probes were used during the eight-week study. Six DIBELS passages were selected, and the Spache readability scores on the probes ranged from 2.6 to 2.7. The probes were arranged into six possible orderings so that each passage appeared in each position. Students were randomly assigned into six groups, and each group received the passages in a different order. The first three passages were read during one sitting in the first week, and the remainder of the probes were read each during week numbers three, five and seven. There were several interesting results in this study. First, the reading fluency scores of the probes were highly correlated (.87 to .93), which suggested that the passages used had high

reliability and validity in assessing oral reading fluency, according to the authors. In terms of the three probes administered during Week One of the study, there were significant effects found for the passage ($p<.0001$), but there were no significant effects for passage order or an interaction between passage and passage order. According to the authors, this finding indicates that the passages cannot be considered equivalent in terms of means and variances. In order to determine if there are similar effects for progress monitoring, scores on the remaining probes that were administered during weeks three, five, and seven were analyzed using an individual growth model. Examination of the slopes of the growth trajectories indicated that they were significantly influenced by the placement of the more difficult stories. The researchers used an equipercentile equating method in order to convert raw WRCM scores and create a conversion table for each probe (Francis, Santi, Barr, Fletcher, Varisco, and Foorman, 2008). There are several practical implications of this study. First, there is evidence that even probes that are purported to be equivalent forms of one another are not necessarily equivalent when used for both benchmarking and progress monitoring. Another important finding is that R-CBM probes can be equated using student performance and equating models.

Some early guidance in terms of data collection methods

Assessing Special Children is one of the first comprehensive publications regarding Curriculum-Based Measurement; it is a book in which many of the pioneers in CBM have written chapters regarding various aspects of CBM (Shinn, 1989). One of the chapters details how to collect R-CBM data. The author says that "…three passages are read…" and the "summary score of interest is the student's median score…as it is the least biased

estimator for small samples," and Hayes (1973) is cited. Shinn (1989) goes on to say that the median score is the best measure of "central tendency,…as it basically ignores extremely low or high scores…and does not involve any calculations." The author explains that during the screening process, "probes are administered for 3 days…within a 5-day period." Those median scores are to be graphed and compared to the average of the student's grade-level score. It is assumed that the author has used this method and found it to be useful, however, the procedures are not reported to be derived from empirical study. There is not a detailed explanation of specifically how to collect the data in this publication, and there is little reported scientific research to explain why one way is better than another, other than some case studies.

There is some research regarding distributed practice effects that can be applied to R-CBM procedures. There is a long history of debate over the advantages and disadvantages of distributed versus massed repetitions (Underwood, 1961). Distributed repetitions refers to repetitions that are distributed over a space of time, while massed repetitions occur within a single time period. According to Melton (1967) there is an advantage for distributed repetitions in terms of learning—this has been called the spacing effect. Glenberg (1976) found that distributed practice consistently resulted in better long-term retention of the skill, while massed practice was superior only if memory was required for a short interval. These concepts refer to learning and memory, however, they could be applied to the administration of R-CBM. There is no known study that examines whether massed or distributed data collection when benchmarking or progress monitoring with R-CBM probes is better or if there are differences in outcomes depending upon which method is chosen.

Shinn (2002) gives more specific explanations of how to administer each reading probe in the AIMSweb Training Workbook: *Administration and Scoring of Reading Curriculum-Based Measurement (R-CBM) for use in General Outcome Measurement.* For the purposes of collecting Benchmark data (usually tri-annual universal screenings at Tier I), it is stated that three probes are used at each of three data collection points: fall, winter, and spring. The median score obtained is the score that is used. In terms of on-going progress monitoring, it is stated in the *AIMSweb Training Workbook: Progress Monitoring Strategies for Writing Individual Goals in General Curriculum and More Frequent Formative Evaluation* that "there is no single formula that has been validated as to how frequently students must be tested" (Shinn, 2002, p. 43).

There are three guiding principles that Shinn (2002) says should be used in determining the frequency of progress monitoring. The first principle is that the frequency of data collection should be related to the severity of the reading difficulties of the student. It is reported that the greater number of data points obtained and the sooner they are obtained, the greater the ability to determine a student's rate of progress and, therefore the need for additional intervention or not. Shinn states that "as a rule, it seems that a minimum of 7-10 data points are necessary to make a reliable decision about student progress," which was determined through empirical study (Shinn & Good, 1989) (Shinn, 2002, p. 43). Shinn (2002) goes on to state that students with more severe reading problems need to be assessed more frequently in order to make changes in interventions as needed. It is stated that it is "desirable to monitor…progress 2 times per week, if feasible" (p. 44). He goes on to say that "there appears to be no benefit to decision making for testing more than 2 times per week." It is also indicated that a minimum number of data points has not been established.

The second principle is the need to balance the ideal scenario with what is feasible for teachers and staff. Although "Best Practice" would advocate progress monitoring to occur 2 times per week, this may not be feasible for all students. Within a problem-solving model, up to 20% of a school's population could be in need of more frequent monitoring, which might still be difficult to attain. It is estimated that it takes 2.5 minutes for testing and scoring for each student with the administration of a single probe, and several planning charts are offered.

The final guiding principle with respect to frequency of administering probes for progress monitoring is that the less frequently a student is monitored, the higher quality the data must be. It is recommended by the author that two to three probes be administered if the assessment frequency is more than once per 2 weeks. The median score of the three scores is used as the data point, because it "increases their likelihood of obtaining a high quality estimate of how their students are doing" (Shinn, 2002, p. 45). This is pertinent to the present argument, because most of what is presented by Shinn (2002) is not based upon empirical study; additional research is needed related to the manner in which the probes are administered and whether the median score is a reliable estimate of performance.

Shapiro (2004) indicates that during the initial stages of CBM, students need to be assessed in order to determine the appropriate instructional placement; that is, the level at which each student reads at an instructional level. A mastery level would indicate that a student reads a particular grade-level probe with adequate proficiency, while a frustrational level indicates the material is too difficult for the student. According to Shapiro, in order to maximize learning rates and prevent ceiling effects, students should receive progress monitoring at their instructional level. In order to determine the instructional level, a student

would be given three probes at his/her grade level, and three probes at the grade levels above and below. The median score would be compared against the local and/or national norms to determine the instructional level. Shapiro does not indicate specifically how the three probes should be administered, which is one focus of the present study. Differences could exist in median WRCM score if administration of the three probes occurs on one day or across three days in the week. As stated before, the mean score is the suggested score to use, however, the mean score has not been examined in terms of its utility in the CBM process.

With respect to progress monitoring, Shapiro (2004) recommends weekly monitoring. It is recommended that one to two probes would be given weekly. There is no direction in terms of how many days need to separate each data collection point. Shapiro (2004) states, "Students are asked to read each passage aloud for one minute…unlike the assessment of instructional placement, [R-CBM] monitoring involves only a 1-minute reading sample" (p. 239). It is not explicitly clear why three probes are needed to control for any potential variability during the assessment of instructional placement and only one probe is needed during progress monitoring. Additional research is needed to investigate whether one probe per week for progress monitoring is a reliable and valid method of measuring student progress (when compared to the use of more probes).

Impact of CBM collection on decision making

Because important decisions are made for children based on their performance on R-CBM probes, it is essential to ensure that the manner in which R-CBM data is collected is practical yet reliable and valid. There has been little research to guide how R-CBM probes are utilized in the classroom to monitor progress. The manner in which data is collected does

not appear to be based on research—it is rather arbitrary and based on conjecture. Interestingly, there have been no studies that examine if there are differences in the way data is collected for weekly progress monitoring. Shinn (2002) stated that no single method is generally validated. Finally, using the median WRCM score has been the accepted practice in CBM. However, in doing this, two thirds of the data collected are discarded—the two other WRCM scores. There has been little research examining whether there may be meaningful differences between the mean and median scores for progress monitoring or eligibility determination, so an investigation into the differences in outcomes if the mean score is used versus the median score is warranted.

Purpose and Significance of the Study

The present study is proposed to determine whether there are differences in outcomes based on the manner in which the CBM probes were collected. A secondary aim of the study is to determine if there are significant differences in R-CBM scores if the median or the mean score (of three R-CBM probes) is utilized. It is not known if there is a significant difference in Words Correct per Minute (WCPM) when R-CBM probes are given on the same day or spread across days in the week. Because RTI is evolving, there are no clear guidelines in terms of how frequently it is necessary to measure performance. Shapiro (2004) and Shinn (2002) state that three probes should be given during assessment of instructional placement and during the benchmarking process, and that the median score of the three probes should be used due to possible variation in probe difficulty and student performance. In terms of progress monitoring, the direction is somewhat vague. It is noted by Shinn (2002) that one to three probes should be given at each data collection period, depending upon the frequency of

data collection. Shapiro states that one to two probes should be given each week. There are differing opinions regarding how and when probes should be given, however, these procedures have not been studied to see if the manner in which the data are collected makes a difference. It is important to know how often and how many R-CBM probes need to be administered in order to reliably and adequately measure student progress, because performance on weekly probes will impact the decision regarding the type of intervention children receive. Because this CBM data can also be used to make high-stakes decisions (i.e., SLD eligibility), it is important for the data collection process to be standardized, reliable and valid in order to help teachers and educational staff engage sound data based decision-making.

Research Questions

1. Is there a significant difference in reading fluency, as indicated by Words Read Correctly per Minute (WRCM) between groups if Curriculum-Based Measurement-Reading (R-CBM) probes are administered in the following manner:
    a. if three probes are administered to students on each of three days per week;
    b. if only one probe is administered to students on each of three days per week;
    c. if three probes are administered to students one day per week;
    d. or if one probe is administered to students one day per week.
2. Is there a significant difference in Words Read Correctly (WRC) if the median or the mean score for each of three conditions is used to indicate reading fluency measured by three Curriculum-Based Measurement-Reading (R-CBM) probes:
    a. in one day of the week (such as comparing the mean and median scores to each other when all three scores are collected on Wednesdays)
    b. in one day of each of three days in the week (such as comparing the mean and median scores to each other when three scores collected on Mondays, Wednesdays, and Fridays)
    c. across three days in the week (such as comparing the mean score of the three probes when one score is collected each on Monday, Wednesday, and Friday to the median score, which is collected on Wednesday)

Hypotheses

Hypothesis I: There will be a difference in median WRCM between the groups.

Hypothesis II: There will be a difference in WRCM between the median and the mean scores in each of the groups.

CHAPTER III

Methods

Participants and Setting

The participating school is a public school located in the south central part of the United States. In January, 2010, the principal investigator and data collector supervisor met with the principal of the school along with the second-grade teachers to invite them to participate in the study. The principal investigator and the data collector supervisor presented an oral presentation to the principal and the second-grade team of six teachers explaining what CBM is and what would be asked of each student and teacher. All of the teachers agreed to participate and signed the consent form. A copy of the outline of the oral presentation to the principal and teachers, as well as the teacher consent form appear in Appendix E and Appendix C.

Participation in the study required that students met two inclusion and two exclusion criteria. The inclusion criteria required that the student be in second grade and reading on second-grade level. Reading level was determined by the student's DRA (Developmental Reading Assessment) level, a reading test used by the teachers to measure progress in reading (Celebration Press, 2001). All of the students in the study were exposed to the general education curriculum and were not removed from their regular classes for more intensive instruction. Students whose reading level is lower than second grade were excluded, because they would not have been able to read the second-grade passages at an instructional level (Shapiro, 2004). When analyzing the data, data was excluded for students who missed any of the data collection sessions (Poncy, Skinner, & Axtell, 2005).

Participants were recruited from the school whose principal and teachers agreed to participate. Students were invited to participate through a recruitment letter sent home. The informed consent document was attached to the recruitment letter. The informed consent letter is a document explaining the purpose of the study and what will be asked of the parents and students; in this letter it was stated that participation in this study was voluntary and that no penalty would be incurred for withdrawing at any time. The recruitment letter and informed consent document were placed in each child's take-home folder and sent home. Parents returned the letter to the teachers via the same folder. The teachers kept the consent forms in a sealed envelope and gave them to the principle investigator. Only the students whose parents returned the signed informed consent document participated in the study. The recruitment letter can be found in Appendix A, and the parental informed consent document can be found in Appendix B.

In February 2010, the primary investigator gave the recruitment letter and parental consent forms to the teachers to send home with the students. There were 162 students in the second grade of the participating school. The parents were asked to return the consent forms within 2 weeks; 54 parents returned the consent form with signatures indicating consent for their child to participate. Data collection began in March 2010.

Prior to collecting data, the data collectors explained to each student that the study was voluntary and that no penalty would be incurred for withdrawing at any time. This information was outlined in the assent document. The assent document was read individually to students prior to data collection. Students were asked to sign the assent document if they agreed to participate. A copy of the assent document is located in Appendix D.

Measures

AIMSweb (Edformation, 2007) and Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (Good & Kaminski, 2002) are companies that develop R-CBM probes commercially; probes from these companies were used for the present study. Probes from both companies needed to be used due to the large number of probes needed (54), and neither company alone produced enough. The second-grade probes were used for all children. Each probe contains approximately 250 words typed in 14-point font on an 8X11.5 inch piece of paper. Because the probes are developed from different companies, a readability analysis was conducted using computer-based software (Micro Power & Light Co., 1996). The first 100 words of each passage were entered into the program to determine if there were differences across the probes. Table 18 shows the readability score for each probe used in the study. The readability scores for the probes ranged from 1.8 to 3.4. The implications of this variation in probe difficulty potentially impacts student performance.

Procedure

Approval from the University of Houston's Committee for the Protection of Human Subjects was obtained prior to data collection. Approval from the school district's Internal Review Board was also obtained prior to the collection of data. The school district also performed background checks on all of the data collectors before the study began.

R-CBM data was collected during a six-week period during the spring semester of 2010. Eight individuals other than the primary investigator collected all data; all individuals were blind to the purpose of the study. They were trained by the primary investigator in standardized administration and scoring procedures described by Shinn & Shinn (2002) and

in the AIMSweb manuals (Edformation, 2007). The training consisted of the data collectors watching videos of students reading passages while the data collectors scored the corresponding probes. The data collectors timed the student on the video for one minute and marked each word read incorrectly. In order to determine the accuracy of scoring, the principal investigator compared the data collectors' scoring response to the correct scoring response. Responses were considered correct if the data collectors' responses matched the scoring key. A scoring response was considered incorrect if the data collectors did not mark it as incorrect when it should have been marked or marked one that should not have been marked as incorrect. Scoring accuracy was calculated by dividing the number of agreements by the number of agreements plus disagreements and multiplying by 100. Training continued until a minimum of 95% accuracy was demonstrated on scoring (Marston, 1989).

As an additional check on scoring integrity, the data collectors participated in a practice session and were evaluated in terms of competency of administration and scoring. The data collectors administered several probes to children of the principal investigator's friends with their parents' permission. Another person, who is well-trained in the administration of R-CBM probes, other than the primary investigator, scored the practice sessions along with the primary investigator to ensure that the data collectors were proficient. Interrater reliability of the practice sessions reached 100% for all of the data collectors. According to Marston (1989), research on the interrater reliability of the administration of CBM probes indicates levels should be higher than .95, so this was the criterion for this study. One of the data collectors volunteered to organize all of the probes and prepare the materials for each testing session. She was considered to be the data collection supervisor. The supervisor scored 4 days (22%) of the eighteen data collection sessions in order to

evaluate for the integrity of standardized procedures. She and the primary investigator randomly chose dates on which the integrity checks would occur. Each of the data collectors was informed that they would be periodically monitored to insure integrity of data collection. The third and fifth weeks were chosen, and it was determined that each data collector would be observed at least once during each of the weeks. The formula used for interobserver agreement was agreements divided by agreements plus disagreements. Interobserver agreement for this study was .955, which implies a high level of data collection integrity.

Teachers provided the researcher with Developmental Reading Levels (DRA levels), which indicate whether or not the students are reading on a second-grade level. Developmental Reading Assessment is a reading test that was developed using criterion-referenced norms that incorporates reading fluency, reading accuracy, and reading comprehension (Celebration Press, 2001). Students who had DRA levels of at least 18, which indicates beginning second-grade reading level, were included in the study, while students who had DRA levels below that were excluded. The initial R-CBM probes were examined to determine if students are reading the second-grade passages at an instructional level based on national norms (Shapiro, 2004). All of the students whose DRA levels were 18 or above were reading within at least an instructional level with second-grade probes.

Each student was individually administered R-CBM probes in an environment as free from distractions as possible. Students joined the data collectors at a desk just outside the classroom door in a quiet hallway. At times, the classes would take restroom breaks, however data collection occurred in the mornings during a time in which the students were in their classrooms. On the rare occasion if a class needed to exit their classroom and occupy

the halls during data collection, the data collector paused testing to minimize disruptions to testing.

Each student was administered three probes on each of three days each week. The data collection usually was on Monday, Wednesday, and Friday, however, due to a spring holiday and a field trip, there were instances in which data collection occurred on a Tuesday instead of Monday or on Thursday instead of Friday. During Week Three, data collection occurred on Thursday, and during Week Four it occurred on Monday due to the holiday. During Week Five, the students took a field trip on Monday, so data collection occurred on Tuesday. Nevertheless, data collection occurred three days each week. To address the research questions, the data was categorized into four contrived groups or "treatment conditions." Each group represents a different method through which probes were collected. The following probes were selected for each of the four treatment conditions. One group (Group A) was administered three probes, three times each week, such as on Monday, Wednesday, and Friday. The second group (Group B) was administered three probes administered one day per week, always on Wednesday. The third group (Group C) was administered the first probe from each of the three days in the week. The fourth group (Group D) was administered only the first probe collected on Wednesdays. This design enabled the study to have four "groups" of data that simulate different data collection methods while minimizing error due to subject differences not associated with the conditions of the study.

Each participant read aloud to the data collector an AIMSweb (Edformation, 2007) passage or a passage from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (Good & Kaminski, 2002) for one minute. Each student read a total of three probes during

each session. Once each participant finished reading the three probes, he/she was given generic verbal praise (such as, "nice work" or "good job") by the data collector. The participant then returned to his/her class and called for the next student to join the data collector.

Each Reading CBM probe was administered only one time throughout the study. Each probe was scored by tallying the number of words read for each prompt and subtracting the number of words read incorrectly. This score yields Words Correct per Minute (WCPM), and this procedure is outlined in Shinn (2002). A word was marked as incorrect if the student omitted a word, mispronounced a word, or could not read the word within the three-second time limit. If the student did not say a word in three seconds, the administrator told the student the word.

Teacher participation

Teachers were not required to collect any data for this study, however their cooperation was an integral part of the data collection. The primary investigator, along with the data collector supervisor, met with the teachers as a group during their conference period prior to the beginning of data collection. The primary investigator gave an overview of Curriculum Based Measurement (CBM) to the teachers and explained its pertinence to the Response to Intervention (RtI) model. The primary investigator explained the procedure for an organized way to administer probes to all of the participants. In exchange for the participation of the teachers the investigator provided the R-CBM data to the teachers after the study was completed and provided training regarding how to collect and interpret R-

CBM data, as well as reading interventions. In addition, a gift card in the amount of $25 was given to each participating teacher.

CHAPTER IV

Results

Analysis

Data analysis occurred in three steps. First the demographic data was examined. Next, the main research questions were addressed using a two-way repeated measures Analysis of Variance; post hoc analysis focused on pairwise comparisons and further examination of the data by week. Finally, additional analyses were conducted to explore student performance on probes; the reading probes were analyzed to see how much each probe varied in terms of readability and to see if student performance was related to probe difficulty. All of the median and mean scores for each condition over the six week study can be found in Appendix G, listed as Tables 1-8. For each condition there is a separate table for the median and mean WRCM scores.

Demographic Information.

Fifty-three students returned signed consent forms, indicating that they could participate in the research study. Four of the students were excluded from the study because their DRA level was below the cut-off score of 18. Seven of the students were excluded from the study because they were absent on data collection days during the study. The final sample used for analysis included 18 boys (42.9%) and 24 girls (57.1%). With regard to race, 9 were African American, 7 were Asian, 10 were Caucasian, and 11 were Hispanic. Please see Table 9 for a breakdown of gender and race.

Table 9

Demographic Information

Gender	African American	Asian	Caucasian	Hispanic
Male	22.2%	27.7%	27.7%	22.2%
Female	41.6%	8.3%	20.8%	29.1%

Research Questions.

The primary focus of the study was to determine if there are differences in WRCM when probes are administered on different days of the week and when the number of probes given is different. There were four conditions that were simulated in this study, and each condition yielded a weekly median score. Condition A was the median score of 9 probes that were administered on 3 days during the week. Condition B was the median score of 3 probes that were administered on 3 different days of the week. Condition C was the median score of 3 probes that were administered all on one day of the week. Condition D was the first probe that was administered on the middle collection day of the week, which simulates typical, weekly progress monitoring where only one probe is administered each week. Table 10 is a graphic representation of how the data was used to form the four conditions. The shaded areas signify which probe was used for the condition.

Table 10

Data Collection Methods for Determining Conditions		
Condition A (median score of all 9 Probes across 3 days)		
Monday	Wednesday	Friday
1st probe	1st probe	1st probe
2nd probe	2nd probe	2nd probe
3rd probe	3rd probe	3rd probe
Condition B (median score of 1st probe on 3 different days)		
Monday	Wednesday	Friday
1st probe	1st probe	1st probe
2nd probe	2nd probe	2nd probe
3rd probe	3rd probe	3rd probe
Condition C (median score of all 3 probes administered on Wednesday)		
Monday	Wednesday	Friday
1st probe	1st probe	1st probe
2nd probe	2nd probe	2nd probe
3rd probe	3rd probe	3rd probe
Condition D (1st probe administered on Wednesday)		
Monday	Wednesday	Friday
1st probe	1st probe	1st probe
2nd probe	2nd probe	2nd probe
3rd probe	3rd probe	3rd probe

Are there differences in students' WRCM between all four conditions?

An a priori power analysis (Cohen, 1988) was conducted in order to determine the number of students that would be needed in order to provide adequate power (.95), a medium effect size (.25), and an alpha level of .05, and the results indicated that a sample size of 28 students would suffice. For this analysis, data from 42 of the students was used, because the statistics software did not allow for any missing values. A two-way repeated measures ANOVA was conducted to evaluate the effect of Testing Condition on WRCM outcomes. The dependent variable was the WRCM score. The within-subjects factors were testing conditions (Conditions A, B, C, and D), and the number of treatment conditions (levels) in this analysis is four (k=4),

and they served as the primary independent variables. Although not one of the research questions, another within-subjects condition was time, and there were 6 levels (weeks) of this independent variable.

Statistical assumptions were assessed to determine any violations before proceeding with the analysis. According to Gravetter and Wallnau (2004) there are several assumptions for using this type of statistic. The data for a repeated measures ANOVA should be measured on an interval or ratio scale; this assumption was met. The data should be normally distributed. The Kolmogorov-Smirnov test (K-S) and Shapiro-Wilk (S-W) test are designed to test normality by comparing data to a normal distribution with the same mean and standard deviation; Skewness and Kurtosis are taken into account simultaneously (Fan, 2010). Normality was assessed using the K-S and S-W tests, and showed no statistically significant scores, indicating no normality violations within the distributions of the dependent measures, except for one variable for each of the tests. For the K-S test of normality, data for Week 6 Condition A was significant ($p = .018$) and for the S-W test, data for Week 3 Condition A was significant ($p = .046$). Further investigation of those variables found that although the K-S and S-W tests were significant, indicating a possible violation of normality, the Skewness and Kurtosis scores were within the recommended cut-off of +3 and -3 (Fan, 2010). In addition, the Normal Q-Q plots did not show major deviation from the normal line, so the data was considered to be normally distributed. Table 11 in Appendix G shows all of the K-S and S-W values.

Another assumption of the repeated measures ANOVA is the independence of observations. This assumption was met, considering each student's performance would not be dependent upon another student's performance. Use of a random sample would also satisfy this assumption, however the sample is not necessarily random, considering students were recruited

from one school and also clustered within classrooms. Using this type of ANOVA requires the assumption of sphericity. According to Field (2009) sphericity is similar to homogeneity of variance, in that it is necessary that each pair of scores has equal variances when examining each pair of treatment levels. Mauchly's Test of Sphericity is used to determine sphericity, and if the assumption is not met, there are several corrections that can be used. When the assumption of sphericity was not met, the correction used for this analysis was the Greenhouse-Geisser correction, which adjusts the degrees of freedom (Field, 2009). An initial analysis using a two-way repeated measures ANOVA revealed that there were differences in weekly WRCM among all conditions when using the weekly median scores. Mauchly's Test of sphericity was significant for testing condition ($p < .001$), indicating that the variances of differences are significantly different; the assumption of sphericity was not met. Because the assumption of sphericity was not met, Greenhouse-Geisser values were used; there was a significant main effect for condition $F(2.32, 95.11) = 22.19$, $p < .001$. This finding implies that there is a difference in outcome (WRCM scores) depending upon the day of the week and the number of probes administered. Although not a research question, the main effect for week was also examined. The assumption of sphericity was met for this analysis ($p = .658$), therefore, no corrections were needed. There was a significant main effect for week $F(5, 205) = 6.28$, $p < .001$. This result suggests that there were differences between WRCM scores between each of the weeks, which is what would be expected. The interaction between Condition and Week was significant, $F(9.03, 370.22) = 16.58$, $p < .001$, which was not initially a research question for this study. This finding suggests that there were differences between the conditions for some weeks, but not for other weeks; depending on which week was examined made a difference on whether the outcomes for the conditions were significantly different from each other. For the interaction effect, Mauchly's

test of Sphericity was significant ($p < .001$), so the Greenhouse-Geisser correction was used. Pairwise comparisons revealed no significant difference ($p = 1.00$) between Condition B, (1 probe on each of three days per week) and Condition C (3 probes administered on one day). This finding indicates that there is no difference in outcome if three probes are administered on one day or three probes spread out over three days during the week over the course of six weeks. In order to determine if there are differences in students' WRCM when one probe is administered versus three, pairwise comparisons were examined. There was a significant difference between Condition B and D ($p < .001$) and Condition C and D ($p = .001$). These results suggest that there is a difference in student outcome when one probe is administered in lieu of three probes over the course of a six-week time period, which contradicts current research and recommended practice according to Shapiro (2004) and Shinn (2002). Table 12 summarizes the findings from the Repeated Measures ANOVA.

Table 12

Two-Way Repeated Measures Analysis of Variance Results for Differences in Students' WRCM Among All Four Conditions

Effect	Mean Square	df	F	p	Partial Eta Squared
Condition	*1056.54	*2.32	22.19	.000	.351
Error	*47.621	*95.11			
Week	1070.404	5	6.280	.000	.133
Error	170.441	205			
Condition*Week	*1106.52	*9.03	16.58	.000	.29
Error	*66.73	*370.22			

*Indicates Greenhouse-Geisser correction value used

Are there differences in students' WRCM between each of the Conditions for each week?

Because there was an interaction effect between Condition and Week, post hoc analyses were conducted in order to determine if there were differences between the conditions for each of the six weeks. Repeated Measures ANOVAS were conducted for each of the six weeks, with Condition being the only independent variable and WRCM being the dependent variable. For all six analyses, Mauchly's test of Sphericity was significant ($p \leq .001$), so the Greenhouse-Geisser correction value was used for all results. An alpha level of .001 was used to determine significance in order to control for experiment-wide alpha (minimize the Type I error). For Week One, there was a significant difference among the four conditions. There was a significant main effect for condition $F(2.31, 104.08) = 11.40$, $p < .001$. Pairwise comparisons indicated that there were no significant differences between Conditions A and Condition B ($p = .125$), Conditions A and C ($p = .002$), Conditions B and C ($p = .004$), and Conditions C and D ($p = .163$), however there were significant differences between Conditions A and D ($p < .001$), and Conditions B and D ($p < .001$). For Week Two, the differences among the conditions were significant $F(2.24, 100.70) = 32.26$, $p < .001$. Pairwise comparisons showed significant differences between four of the conditions during Week Two: Conditions A and C ($p < .001$), Conditions A and D ($p < .001$), Conditions B and D ($p < .001$) and Conditions C and D ($p < .001$). The differences between Conditions A and B ($p = .007$) and Conditions B and C ($p = .011$) were not significant. Week Three's data indicated significant differences among conditions for the ANOVA statistic $F(2.35, 101.17) = 17.48$, $p < .001$. Conditions A and B ($p = .020$), Conditions A and C ($p = .013$), and Conditions B and C ($p = .907$) were not significantly different from one another in terms of pairwise comparisons, however the remaining conditions were significantly different from each other: Conditions A and D ($p < .001$), Conditions B and D

($p < .001$), and Conditions C and D ($p < .001$). For Week Four, there was a significant difference among the four conditions. There was a significant main effect for condition $F(2.27, 100.07) = 7.94$, $p < .001$. Pairwise comparisons indicated that there were no significant differences between Conditions A and B ($p = .799$), Conditions A and C ($p = .002$), Conditions A and D ($p = .003$), Conditions B and C ($p = .007$), and Conditions C and D ($p = .404$), while there was a significant difference between Conditions B and D ($p = .001$). There was a significant difference among all conditions for Week Five $F(2.33, 104.87) = 15.89$, $p < .001$, and pairwise comparisons found significant differences between Conditions A and B ($p < .001$), Conditions A and D ($p < .001$), and Conditions C and D ($p < .001$). There was no significant difference between Conditions A and C ($p = .015$), Conditions B and C ($p = .100$) and Conditions B and D ($p = .005$). For the final week of the study, there was a significant difference among all of the conditions $F (2.17, 95.60) = 24.80$, $p < .001$. Pairwise comparisons revealed no significant differences between Conditions A and B ($p = .003$), Conditions A and D ($p = .002$), and Conditions C and D ($p = .253$); there were significant differences for the other three condition pairs: Conditions A and C ($p < .001$), Conditions B and C ($p < .001$) and Conditions B and D ($p < .001$). These variable results based on the week the data was collected calls the reliability of the WRCM scores into question. Please see Tables 13 and 14 for a summary of the ANOVA results and Pairwise Comparisons. See Table 15 for a comparison of significant findings of Repeated Measures ANOVAs for each of the six weeks. Graph 1 shows the differences in mean WRCM for each condition by week.

Table 13

Repeated Measures Analyses of Variance Results for Differences in Students' WRCM Among All Four Conditions by Week

Effect	Mean Square	df	F	p	Partial Eta Squared
Condition Week 1	615.16	2.31	11.40	.000	.20
Error	59.95	104.08			
Condition Week 2	2067.77	2.24	32.26	.000	.42
Error	64.10	100.70			
Condition Week 3	1000.96	2.35	17.48	.000	.29
Error	57.27	101.17			
Condition Week 4	293.53	2.27	7.94	.000	.15
Error	36.96	100.07			
Condition Week 5	659.05	2.33	15.89	.000	.26
Error	41.48	104.87			
Condition Week 6	1326.10	2.17	24.80	.000	.36
Error	53.47	95.60			

Greenhouse-Geisser correction value used for all ANOVAS

Table 14

Pairwise Comparisons of Each Condition by Week

	Condition Pair	Significance
Week 1		
	Conditions A and B	.125
	Conditions A and C	.002
	Conditions A and D	.000
	Conditions B and C	.004
	Conditions B and D	.000
	Conditions C and D	.163
Week 2		
	Conditions A and B	.007
	Conditions A and C	.000
	Conditions A and D	.000
	Conditions B and C	.011
	Conditions B and D	.000
	Conditions C and D	.000
Week 3		
	Conditions A and B	.020
	Conditions A and C	.013
	Conditions A and D	.000
	Conditions B and C	.907
	Conditions B and D	.000
	Conditions C and D	.000
Week 4		
	Conditions A and B	.799
	Conditions A and C	.002
	Conditions A and D	.003
	Conditions B and C	.007
	Conditions B and D	.001
	Conditions C and D	.404
Week 5		
	Conditions A and B	.000
	Conditions A and C	.015
	Conditions A and D	.000
	Conditions B and C	.100
	Conditions B and D	.005
	Conditions C and D	.000
Week 6		
	Conditions A and B	.003
	Conditions A and C	.000
	Conditions A and D	.002
	Conditions B and C	.000
	Conditions B and D	.000
	Conditions C and D	.253

Table 15

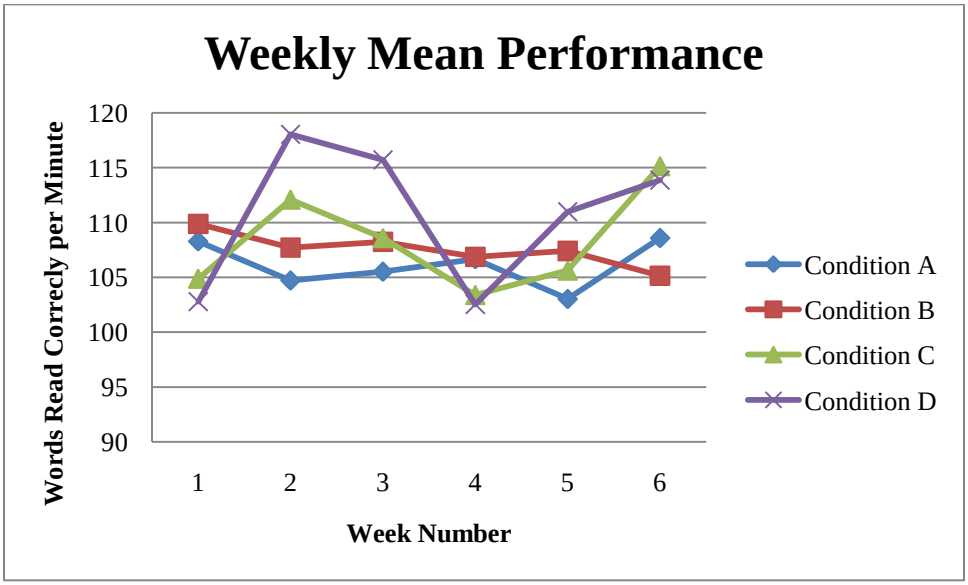
Comparison of Significant Findings of Repeated Measures ANOVAs for Each of the Six Weeks

Condition Pair	Significant	Not Significant	
Conditions A and B	1	5	Similar
Conditions A and C	2	4	Similar
Conditions A and D	4	2	Different
Conditions B and C	1	5	Similar
Conditions B and D	5	1	Different
Conditions C and D	3	3	X

Graph 1 shows the differences in mean WRCM for each condition by week. It illustrates that Condition D shows a greater degree of variability than the other conditions. Conditions A and B appeared to be more similar to each other; additionally, these two conditions seem to show less variability. It appears that there is less variability when three or more probes are used and administered on the same day.

Graph 1

Weekly Mean Performance for All Four Conditions



Is there a difference when the weekly Mean score is used instead of the weekly Median score?

The second research question asks if there are differences if the mean or median score is used, and the data was analyzed for three of the four administration conditions of the study (Condition D was only one score, so there could be no comparison). Analysis for the second research question was done using two-way repeated measures ANOVAs to compare the median to mean score across weeks in each of the three conditions. An alpha level of .001 was used for this analysis. For all three conditions, there were no significant differences between the median and mean scores. There were two independent variables: Mean or Median Condition and Week. There were two levels for Condition and six levels for week. The dependent variable was the WRCM score. Although not a research question, the interaction effect for Condition and Week was not significant for all three of the conditions. The main effect for each condition is reported in Tables 16, 17, and 18, however the results for week will not be discussed, because they are not pertinent to this study. For the main effect of Condition A (the median and mean scores of all 9 weekly probes), Mauchly's Test of Sphericity was not significant, so sphericity was assumed. There was not a significant main effect $F(1, 44) = .01$, $p = .944$. For Condition B (the median and mean scores of the first probe given on each of 3 days of the week) Mauchly's Test of Sphericity was not significant, so sphericity was assumed. There was not a significant main effect $F(1, 44) = 1.84$, $p = .182$. For Condition C (the median and mean scores of the three probes given on Wednesdays) Mauchly's Test of Sphericity was not significant, indicating no violation of sphericity. There was no significant main effect $F(1, 5) = .174$, $p = .679$. This finding suggests that over time, there is no difference in outcome if the median or mean score is used.

Table 16

Two-Way Repeated Measures Analysis of Variance Results for Differences in Students' WRCM Between Median and Mean WRCM scores Across Condition A

Effect	Mean Square	df	F	p	Partial Eta Squared
Median vs. Mean	.092	1	.01	.944	.00
Error	18.42	44			
Week	*602.88	*4	6.28	.000	.13
Error	*95.95	*176.01			
Median vs. Mean * Week	*122.64	*1.61	2.293	.119	.05
Error	*53.50	*70.73			

*Greenhouse-Geisser correction value used

Table 17

Two-Way Repeated Measures Analysis of Variance Results for Differences in Students' WRCM Between Median and Mean WRCM scores Across Condition B

Median vs. Mean	22.41	1	1.84	.182	.40
Error	12.16	44			
Week	276.31	5	2.47	.033	.05
Error	111.76	220			
Median vs. Mean * Week	*28.89	*4.02	2.27	.064	.05
Error	*12.76	*176.82			

*Greenhouse-Geisser correction value used

Table 18

Two-Way Repeated Measures Analysis of Variance Results for Differences in Students' WRCM Between Median and Mean WRCM scores Across Condition C

Median vs. Mean	18.03	1	.17	.679	.00
Error	103.62	41			
Week	*2971.80	*2.52	6.21	.001	.13
Error	*478.55	*103.49			
Median vs. Mean * Week	*637.23	*1.13	1.37	.252	.03
Error	*463.75	*46.45			

*Greenhouse-Geisser correction value used

CHAPTER V

Discussion

Differences in administration of CBM probes

The results from this study suggest that the day of the week and number of probes administered will affect the WRCM score obtained. The median WRCM score from nine weekly probes was different from the median WRCM score from three weekly probes and from one WRCM score from one probe administered weekly. The literature suggests that the reliability and SEM of WRCM scores improve the as the number of probes administered increases. The previous research coupled with the current differences found suggests that one weekly probe may not be reliable enough to make important decisions. This is concerning, because WRCM scores are used to make important low-stakes and high-stakes decisions, such as: tier placement; the amount and intensity of intervention students receive; if students are responding to interventions and if interventions need to be changed or added; and in some cases, if students qualify to receive Special Education services.

The pioneers of CBM proposed that three probes should be used for universal screenings, and one probe should be used for weekly progress monitoring, however this recommendation is not based upon scientific inquiry. Whether the three probes used for universal screening should be administered on the same day or spread out over three days is also not clear. The current study found that there is no difference if the three probes are given on the same day or spread out over the week. This is important to the practice of R-CBM, because it is most likely easier for educators to administer all three probes during the same session with each student. Until now, there has been no research to show if there is a difference in outcome. Because there is not a

significant difference, educators can choose to administer the probes on the same day or across three days in the week.

Previous research suggests that the more probes that are administered, the more reliable the WRCM score is, and the lower the SEM (Poncy, Skinner, and Axtell, 2005). The current study compared the WRCM outcomes given different administration conditions, such as 9 probes given weekly (3 across 3 days), 3 probes weekly (3 given on one day and 3 given across 3 days), and one probe weekly. The WRCM outcomes were significantly different depending upon the administration condition. This study did not explore the reliability and SEM of the given conditions, so it is not possible to say which condition is more reliable. However, the fact that each condition yielded a significantly different outcome is alarming, and has implications regarding the utility of CBM for high-stakes decision making. If the WRCM score changes due to data collection protocols, the basis for CBM is questionable, because we are not detecting "small but important" changes in reading ability, we are detecting changes in data collection protocols, probe variability, and other unknown factors.

Typically, when conducting a survey-level assessment, three probes are administered, and the median score is used. The same procedure is used during the benchmarking process/universal screening that occurs three times per year to screen students for possible reading difficulties. Three probes are used for assessment for instructional placement and universal screenings, however only one probe is used for progress monitoring. The present study compared the WRCM score when one probe was administered weekly to two other administration conditions where three probes were administered weekly. All yielded significantly different WRCM scores. This finding is similar to the finding from the Poncy, Skinner, and Axtell, (2005) study that found that one probe administered yielded different results

than more probes. This finding is problematic, because it demonstrates, again, that another factor in the student's WRCM score is the data collection protocol that is employed, which means that we are less confident about the meaning of the WRCM score obtained. We cannot be sure whether we are really measuring the student's progress; the WRCM score is dependent upon a number of factors, one being when and the number of probes that are administered. If we cannot be confident that the CBM probe is measuring the student's reading ability, then we are unable to use the WRCM score to make decisions for the student regarding the student's response to intervention and the need for more or less intensive interventions. If the CBM probes are not reliable and valid enough to make day-to-day instructional decisions, they clearly are not reliable and valid for use in Special Education placement decisions.

The question regarding the use of the median score or the mean score is one that has not been explored in depth. The median score is traditionally used to help decrease error associated with variability (Shapiro, 2004). However, the mean score takes into account the student's entire performance. In the present study, there was no difference over time between the median and mean scores. One implication for this finding is that mean scores could be considered, because there is no difference in outcome from median WRCM scores when the mean WRCM score is used over time.

Limitations to the study

There are a number of limitations to this study. The major limitation is the assumption of probe equivalence, which may be the largest threat to internal validity. Differences in WRCM scores among and between subjects could be attributed to effects of the independent variables, but they could also be due to variability in passage difficulty. One method that was used to

attempt to address probe variability was to determine the extent to which the passages are equivalent via the use of a computer program that uses readability formulas. As noted in the review of literature, there exists some debate regarding the effectiveness of this method in establishing probe equivalence.

Because of the large number of probes needed for the study, probes from two different companies were used. In order to determine if there were significant differences between the two groups of probes, an independent samples t-test was performed on the Spache readability scores. There was a significant difference in difficulty between the groups $t(52) = -4.483$, $p < .001$. The mean readability score of the AIMSweb probes was 2.44, while the mean readability score of the DIBLES probes was 2.78. The difference in difficulty between the types of probes used likely had an effect on the WRCM change over time. Student performance was moderately correlated with probe difficulty, therefore, it is likely that the group of more difficult probes being used for the second half of the study had an impact on the growth rate of the students. This finding demonstrates that DIBELS and AIMSweb probes may not be used interchangeably.

According to AIMSweb national aggregate norm tables (aimsweb.com) the average rate of growth for students is 1.2 WRCM per week. For the current study, the median weekly score of all 9 probes administered each week for each student was determined in order to view the performance of each student during the six-week study and examine trends. The total amount of change in WRCM performance was calculated for the entire six weeks. Forty-one percent (N=19) of the students increased in their median WRCM score, while 59% of the students (N=27) decreased in their median WRCM scores. Because different types of commercially-developed probes were used during the study, the data were broken down in terms of the type of probe used. AIMSweb probes were used during the first three weeks of the study, and DIBELS

probes were used during the latter three weeks of the study. With regard to student performance on the AIMSweb probes, 15 students (32%) increased median WRCM from week 1 to week 3, 2 students (5%) had the same median WRCM, and 29 students (63%) decreased median WRCM from week 1 to week 3. DIBELS probes were also evaluated in terms of student performance, and on the DIBELS probes, 26 students (56%) increased median WRCM from week 4 to week 6, 1 student (3%) had no change in median WRCM, and 19 students (41%) median WRCM decreased. Table 19 summarizes this data and can be found in Appendix G. Perhaps measuring WRCM growth over time using growth trajectories and estimates of linear growth rates (such as the data that is provided when using the AIMSweb on-line software and the recommendations of Francis et al. (2008) would provide a more accurate picture of student performance.

With regard to practice effects, it appears that there is no difference between distributed or massed practice. One important consideration for this study, however is the effect on the students of being administered so many probes weekly. One could argue that the students performed differently than what would be expected in a practical situation. Students could have benefitted from practice effects, which could have affected the results. Conversely, they could have suffered from fatigue from taking so many tests. Regardless, as with many experiments in applied settings, results should be interpreted with caution with regard to generalization.

The school in which the study was conducted has an ethnically and racially diverse population of students who attend, and the students who participated reflect that diversity. Even so, a limitation of the study, as with many studies is the generalizability of the data to other populations. The small number of students who participated is a limitation, and a larger population of students could possibly provide greater power to the results which may have detected additional differences. Additionally, the sample only included second-grade students

who were able to read second-grade material, so results are not to be generalized to other grade levels or to students who have been diagnosed with a Specific Learning Disability in reading. Typically, students who would participate in progress monitoring are those who are reading below the 25th percentile in terms of WRCM when compared to peers. Those students' performance would likely be different from the students in the current study, because the students in the current study did not fall below the 25th percentile when compared to the national aggregate norms. This study could be replicated and extended to include multiple schools, years, grades, and students in special education/tiers II and III in order to evaluate the external validity of the findings.

Another limitation of the study is that subjects were assessed by different data collectors at each session. This could have resulted in an increase in error/variability in scores, however due to the number of probes that were given and the number of students, it was not feasible for one person to collect all of the data. In order to attempt to reduce this kind of variability, all of the data collectors were well-trained in the procedures and a high level of proficiency was obtained prior to data collection. Some of the sessions (22%) were observed by two data collectors (one of them being the data collector supervisor) to determine fidelity of data collection procedures. Interrater reliability of the observed sessions was .965.

The current study involved researchers collecting data in the same environment at approximately the same time of day (i.e. during a 2-3 hour block of time in the morning) for each data collection session. The setting remained the same every day, however there were three days in which data collection was not on a Monday or a Friday. Due to Spring holidays, two of the Monday data collection days were moved to Tuesdays; this occurred during Week Four and Week Five. During Week Three, data collection occurred on Tuesday instead of Monday. In

addition, after Week One of the study, there was one full week of spring break, in which the students were on vacation from school. In applied settings, reading probes may be administered under less ideal situations and may not occur in a consistent manner (i.e. same time of day). It is possible that these differences introduce variability from unaccounted sources which could increase error and external validity.

Directions for future research

A great deal of research regarding the utility and psychometric qualities of CBM is underway, and as evidenced by the current study, a great deal more is needed. Because probe equivalence is questionable in the current study, there needs to be more research and development of equivalent forms of probes. It is not possible to determine with certainty student performance and progress unless the probes are more equivalent. Perhaps more studies are needed that are similar to the Poncy, Skinner, and Axtell (2005) study in which reliability of the probes was determined by using test/re-test procedures and the Francis et al. (2008) study that used a raw score conversion to equate probes. This study could also be extended to include select samples of individual groups, such as students with reading difficulties and students who are not proficient in English.

Supplemental Analyses

Several other hypotheses were generated by examining the data from the current study that have practical implications. Because probe variability was a limitation of the study, the R-CBM probes were examined. The range of performance on probes was analyzed, and correlations were performed to determine if there was a relationship between readability scores

and student performance. A correlation was also conducted in order to assess the relationship between student DRA scores and performance on R-CBM probes. Finally trends in student performance were examined and discussed.

Range of performance on each probe and probe correlations

Because there appeared to be a great deal of variability in terms of performance on probes, student performance on each probe was examined, and the high score and low score on each probe was obtained. The highest score on any probe was 193 WRCM, and it occurred during the first day of the study; it was the final probe given for that day. The lowest score on any probe was 44 WRCM, and it occurred during the 5th week of data collection; it was the second probe administered the second data collection day of the week. The highest range in terms of student performance on a probe was 130 WRCM; the lowest score on that probe was 63, while the highest score was 193. That probe was administered on the first day of testing, and was the last probe administered that day. The lowest range of scores on a probe was 78; the lowest score obtained on that probe was 57 and the highest was 135. That probe was administered second, on the final day of the study. Table 20 shows the low scores, high scores, and range of each probe. Each probe is named based on administration week, day, and order. For example, probe 2.3.1 signifies that it was administered week 2 of the study, on the third administration day of that week, and it was the first probe administered that day (week.day.probe#).

It is expected that there would be a range of scores obtained on any given test. Thus, the range of obtained WRCM scores on the individual probes is not necessarily alarming; that scores varied so much within each student is of concern. There were several instances in which a

student's WRCM varied by 20 or more points on the same day. This is similar to a finding by Francis et al. (2008) which found differences of 26 WRCM depending on form choice. This variability is not due to changes in ability, but rather due to other sources of error, such as probe difficulty, interest in the story, and a number of other possible factors reported by Derr and Shapiro (1989), such as evaluator, location of testing, and whether the student is aware of being timed. Using the median score of the probes decreased the high range and variability, which provides support for this practice.

Relationship between readability and student performance

Because it was determined that there was a great deal of variability among the probes, it was hypothesized that probe difficulty contributed to student performance on the probes. Readability scores were determined for each probe using the Readability Calculations program from Micro Power & Light Co (1996). The Spache formula (revised version) computes a score which represents the appropriate grade level for the evaluated material; this formula is designed to assess reading materials from primary to third-grade level. For example, a score of 2.1 would signify that the reading material is beginning second-grade level; a score of 2.9 would signify end-of-year second-grade reading material. The first 100 words of each probe were entered into the program. The program uses its own word list to determine the potentially difficult words and also factors in the total number of words in each sentence.

In terms of readability scores, the lowest score computed for a probe used in the study was 1.8, and the highest score was 3.4. Of the 54 probes used, 1 scored in the first-grade level in terms of readability, 45 scored in the second-grade level, and 8 scored in the third-grade level.

A Pearson correlation was performed in order to determine if there was a relationship between the readability score and the median student performance for each probe. There was a significant inverted relationship between the readability level and the mean student performance. The correlation was moderate ($p = -.310$), which is significant at the 0.05 level. This negative correlation shows that as the readability score of the probe increased in grade-level, there tended to be lower WRCM scores on the probe.

A correlation was also used to determine if there was a relationship between the readability score and the mean student performance for each probe. There was also significant inverted relationship ($p = -.316$), which is significant at the 0.05 level. This finding shows, just as with the median score, as the readability score increased, the mean student performance on the probe decreased. Table 21 summarizes the relationship between readability and median/mean performance.

There is some debate regarding how to accurately measure the difficulty of probes (Poncy, Skinner, and Axtell, 2005), however one method of quantifying the difficulty is through using readability formulas. In this study, the Revised Spache formula was used to determine the grade-level for each probe. According to the developers of the commercial probes used in the study, all of the probes are considered to be alternate forms of one another, which assumes that each probe is equivalent to each other. However, according to the Spache calculations, there was variation in difficulty from late first-grade level to early third-grade level. Thus, one would expect to see differences in performance due to probe difficulty. This is pertinent to the current study, because when measuring student performance, it is expected that we are measuring student growth or the lack thereof, and not the difficulty of the probe. If the probes differ in

terms of difficulty, we cannot be certain that we are measuring growth, because this variability in the probes contributes to measurement error.

There was a moderate correlation between student performance on all of the probes and the readability score of each probe. When addressing trends in the data graphically, it appeared that WRCM scores in Condition D were more variable, therefore a correlation between mean probe performance and readability scores was conducted. A moderate correlation between performance and readability was found. Generally, as the difficulty of the probe decreased, the students were able to read more words correctly per minute. As the difficulty of the probe increased the students read fewer words correctly per minute. Again, this variation in probe difficulty can potentially pose problems, because student performance is related to the probe. One of the foundations of CBM is that student response to intervention can be detected by small changes in student performance on probes. However, the relationship between performance and probe difficulty shows that WRCM scores are not exclusive of other sources of error. The implication of this uncertainty is that we may be using data that is flawed when we make decisions regarding the need for additional instruction or change of intervention during progress monitoring. If we use this data to make decisions about SLD identification in lieu of other data, we could be either failing to qualify students for special education services who need it or placing students in special education who do not need it, thus violating least restrictive environment.

Relationship between DRA and Universal Screening/Survey-level assessment Median score

Performance on R-CBM probes is often compared to other standardized tests, so it was hypothesized that performance on R-CBM probes would be related to scores on the DRA. The

DRA level of each student was provided to the researchers prior to the first day of data collection. The DRA levels were determined (at most) two weeks prior to the commencement of the study. During the first day of data collection, each student was administered three probes, which simulates what would occur during a universal screening and a survey-level assessment to determine instructional level. The median score for each student was correlated with the DRA level score for each student using a Pearson Correlation. A significant relationship was found between the two scores (.546), which is significant at the 0.001 level. The results suggest that DRA levels and performance on R-CBM probes are moderately correlated. See Table 22 for a summary of the relationship between DRA levels and universal screening/survey-level assessment median scores.

When benchmarking and when performing survey-level assessment, three probes are typically administered at the student's expected grade-level in order to determine if the student is reading grade-level material at an instructional level. The median score is used and compared with local and/or national norms (Shapiro 2004). Developmental Reading Assessment (DRA) is a criterion-referenced test that is thought to measure reading abilities, including fluency, accuracy, and comprehension. The positive correlation between the DRA scores and the median WRCM score on the first day of testing for the study was significant, but only moderate in strength. The research suggests that the correlation between reading fluency and reading comprehension is higher (.60 and higher) than the correlation in the current study (.55) (Deno, Mirkin, & Chiang, 1982; Hamilton & Shinn, 2003; Hintze, Callahan, Matthews, Williams, & Tobin, 2002; Shapiro, Edwards, Lutz, & Keller, 2004). This finding provides some support for one of the principles of R-CBM that reading fluency is an indicator and is related to other reading abilities as measured by common testing practices. The current findings, however,

suggest that there is additional variance that is unaccounted for because this moderate correlation only explains a small percentage of the shared variance (30%). This finding is important to the current study, because it highlights the fact that there are additional factors which contribute to WRCM scores obtained by students.

Analyzing trends in the data

Because the WRCM scores appeared to be variable, it was important to systematically examine the data. In terms of addressing trends in the data, the weekly median scores were examined. It was expected that the students' WRCM scores would increase over the course of six weeks, considering an average weekly rate of improvement for a second-grade student is about 1.2 words (according to AIMSweb aggregate norm charts) or 1.5 words (according to Shinn 2002). However, many of the scores did not increase after six weeks. There are several reasons this could have happened, but many of the explanations are speculative. One possibility is that attendance was not consistent during this time due to several spring holidays. There was an entire week when the students were on Spring Break, a 4-day holiday around Easter, and another Friday in which the students went on a field trip. It was hypothesized that perhaps teacher instruction was different and could have been more effective for some teachers than others.

In an effort to determine if there were any differences in student performance due to differences in teacher instruction, the weekly median scores of all 9 probes were aggregated by teacher. Table 20 shows weekly student scores by teacher and the amount of change of each student in the class. Teacher A had 6 students who participated in the study; when looking at change across the 6-week study, 5 of them showed improvement in median WRCM, while 1

showed a decrease in median WRCM. When looking at the change in median WRCM during the first 3 weeks of the study using the AIMSweb probes, 3 students showed an increase, while 3 students showed a decrease in median WRCM. For the DIBELS probes, 5 students increased in median WRCM, while 1 student's median WRCM scores decreased. Nine of Teacher B's students participated in the study, and 5 of them increased in median WRCM, while 4 of them decreased from week 1 to week 6. For the AIMSweb probes, all of the students showed a decrease in median WRCM, while 5 students showed an increase and 4 students showed a decrease with the DIBELS probes. There were 14 students in Teacher C's class who participated; six of them improved in their total median WRCM scores, while 8 of them read fewer words by week 6. For the AIMSweb probes used week 1-3, 7 of the students' median WRCM scores increased while 6 scores decreased. For weeks 4-6, when DIBELS probes were used, 5 students showed increases in median WRCM scores, while 9 showed decreases. Seven students from Teacher D's class were participants, and in terms of total change during the six-week study, 2 students' median WRCM scores increased, one remained the same, and 4 decreased. For the AIMSweb probes, one student's scores showed no change, and the remainder of them showed a decrease in median WRCM scores. For the DIBELS probes, 4 students' median scores increased, one stayed the same, and two decreased. Teacher E had 4 students that participated; one student's median WRCM score increased, while the remainder of them decreased. The trend for the AIMSweb probes was the same. Two students had increased median WRCM scores with the DIBELS probes, and two had decreased scores. The final teacher, Teacher F, had 6 students who participated in the study. Four of her students showed an increase in median WRCM scores from week 1 to week 6. On the AIMSweb probes, 4 of the

students' scores increased, one stayed the same, and one decreased. On the DIBELS probes used during weeks 3-6, 4 students' scores increased and 2 decreased.

When the data was aggregated by teacher, all of the teachers had some students who improved and others who showed regression (see Table 23). It is assumed that all of the students were exposed to the same curriculum, because according to the team leader teacher, all of the teachers teach the same concepts and engage in the same learning and practice activities. None of the students in the study received additional reading instruction, because they were all considered to be reading on second-grade level in terms of reading abilities. The variable performance on the probes from week to week in this study highlights the need for use of a method in which to interpret R-CBM data; simply looking at an increase or a decrease in scores does not reveal much about student progress. Shinn (2002) and Shapiro (2004) report that trends in performance must be analyzed. In particular, slope or rate of improvement should be calculated, and each student's performance should be analyzed by a team on a regular basis. Francis et.al. (2008) recommend using growth trajectories and raw score conversions. Christ and Silberglit (2007) indicate that learning trajectory should be assessed only after 7-10 probes are administered, because as the number of data points increases, the effects of measurement error on the trend line decreases. Shinn and Goode (1989) recommend suspending analysis until at least 6-9 probes have been administered, particularly if the data appears variable. The results of this study, coupled with current trends in research suggest that many data points are needed to decrease error, and decisions should not be made until trend lines are stable.

References

Ardoin, S.P. & Christ, T. (2008). Evaluating curriculum based measurement slope estimates using data from tri-annual universal screenings. *School Psychology Review*, 37, 109-125.Christ & Ardoin, 2009

Beaver, J.M. (2001). *DRA Developmental reading assessment :K-3 teacher resource guide*. Parsippany, N.J.: Celebration Press / Pearson Learning Group

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

Christ, T. J., & Silberglitt, B. (2007). Curriculum-based measurement of oral reading fluency: The standard error of measurement. *School Psychology Review, 36,* 130-146.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. (2nd ed.) Hillsdale, NJ: Erlbaum

Daly, E. J., III, Witt, J. C., Martens, B. K., & Dool, E. J. (1997). A model for conducting a functional analysis of academic performance problems. *School Psychology Review, 26*, 554–574.

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

Deno, S.L. (1989). Curriculum-based measurement and alternative special education services: A fundamental and direct relationship. In M.R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children*. New York: Guilford Press.

Deno, S.L. (2002). Problem solving as best practices. In A. Thomas & J. Grimes (Eds.), *Best Practices in school psychology IV* (pp. 37-56). Bethesda, MD: National

Association of School Psychologists.

Deno, S.L., Marston, D., Shinn, M.R., & Tindal, G. (1983). Oral reading fluency: A simple datum for scaling reading disability. *Topics in Learning and Learning Disabilities*, *2*(4), 53–59.

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

Derr, T.F., & Shapiro, E.S. (1989). A behavioral evaluation of curriculum-based assessment of reading. *Journal of Psychoeducational Assessment, 7*, 148-160.

Donovan, M. S., & Cross, C. T. (2002). *Minority students in special and gifted education.* Washington, DC: National Academy Press.

Dunn, L. M., & Markwardt, F. C., Jr. (1970). Peabody Individual Achievement Test. Circle Pines, MN: American Guidance.

Dykeman, B. F. (2006). Alternative strategies in assessing special education needs. *Education,* 127, (2), 265-273.

Edformation, Inc. (2007). AIMSweb Reading Fluency Probes. Retrieved July 20, 2009, from: http://www.aimsweb.com

Fan, W. (2010). Tests of Normality [Class handout]. Department of Educational Psychology, University of Houston, Houston, Texas.

Field, A.P. (2009). Discovering statistics using SPSS: and sex and drugs and rock 'n' roll (3rd Edition). London: Sage.

Flanagan, D.P., Ortiz. S.O., & Alfonso,V.C. (2007). *Essentials of Cross Battery Assessment, 2nd Ed*. New York: John Wiley & Sons.

Fiorello, C.A., Hale, J.B., & Snyder, L.E. (2006). Cognitive hypothesis testing and response

to intervention for children with reading problems. *Psychology in the Schools, 43,* 835-853.

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(3),* 315-342.

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

Fuchs, L. S., & Richs, D. (1986). Effects of systematic formative evaluation on student achievement: A metaanalysis. *Exceptional Children. 53,* 199-208.

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

Fuchs, L. S. & Fuchs, D. (2001). Responsiveness to intervention: A blueprint for practitioners, policymakers, and parents. *Teaching Exceptional Children, 38 (1),* 57-61.

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

Fuchs, L. S., Fuchs, D., & Hamlett, C.L. (1989) Effects of instrumental use of Curriculum-Based Measurement to enhance instructional programs. *Remedial and Special Education, 10 (2),* 43-52.

Fuchs, L. S., Fuchs, D.. & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education. 9.* 20-28.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior 15*, 1-15.

Glover, T. A., & Diperna, J. C. (2007). Service delivery for response to intervention: Core components and directions for future research. *School Psychology Review, 36*(4), 526 540.

Good, R. H. & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.), Retrieved July 20, 2009 from Oregon University, Institute for the Development of Educational Achievement, http://dibels.uoregon.edu

Gravetter, F.J.; Wallnau, L.B. (2004). Statistics for the Behavioral Sciences 6th edition. Thomson Wadsworth.

Hamilton, C. R., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgements of reading comprehension and oral reading skills. *School Psychology Review, 32*, 228–240.

Hayes, W. L. (1973). Statistics for the Social Sciences. Second Edition. New York: Holt, Rinehart and Winston, Inc.,

Hintze, J. M., Callahan III, J. E., Matthews, W. J., S. Williams, S. A., & Tobin, K. G. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review, 31(*4), 540-554.

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

Hintze, J.M., Shapiro, E.S., & Lutz, J.G. (1994). The effects of curriculum on the sensitivity of curriculum-based measurement in reading. *Journal of Special Education*, 28, 189-202.

Holdnack, J. A., Weiss, L. G. (2006). IDEA 2004: Anticipated implications for clinical practice-integrating assessment and intervention. *Psychology in the Schools*, 43, (8), 871-882.

Howe, K. B., & Shinn, M. M. (2002). Standard reading assessment passages for use in general outcome measures: A manual describing development and technical features. Eden Prairie, MN: Edformation, Inc.

Karlsen, B., Madden, R., Gardner, E.F. (1975) Stanford diagnostic reading test. New York: Harcourt Brace Javonovich.

Kavale, K., Kaufman, A. S., Naglieri, J. A., Hale, J. B. (2005). Changing procedures for identifying learning disabilities: The danger of poorly supported ideas. *The School Psychologist*, 59 (1), 17-25.

Marston, D.B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M.R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp.18-78). New York: Guilford Press.

Melton, A. W. (1967). Repetition and retrieval from memory. *Science 158*, 532.

Merhens, W.A., & Clarizio, H.F. (1993). Curriculum-based measurement: Conceptual and psychometric considerations. *Psychology in the Schools*, 30, 241-254.

Micro Power & Light Co. (1996). Readability Calculations for Windows, from www.micropowerandlight.com

Mirkin, P., Deno, S. L., Tindal, G., & Kuehnle, K. (1982). Frequency of measurement and

data utilization as factors in standardized behavioral assessment of academic skill. *Journal of Behavioral Assessment, 4*(4), 361-370.

Ofiesh, N. (2006). Response to intervention and the identification of specific learning disabilities: Why we need comprehensive evaluations as part of the process. *Psychology in the Schools*, 43, (8), 883-888.

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. *Journal of Psychoeducational Assessment*, 23, 326-338.

Powell-Smith, K. A. (2004, February). *Individual differences in FCAT performance: A national context for our results*. Paper presented at the annual meeting of the Pacific Coast Research Conference, Coronado, CA.

Readability Calculations (1999). Micro Power & Light Co. Dallas: TX

Reynolds, C. R., & Shaywitz, S. E. (2009). Response to Intervention: Ready or not? Or, from wait-to-fail to watch-them-fail. *School Psychology Quarterly*, *24*(2) 130-145.

Shapiro, E.S., & Manz, P.H. (2004). *Academic Skills Problems Direct Assessment and Intervention.* New York: Guilford Press.

Shapiro, E.S., Edwards, L., Lutz, J.G., & Keller, M. (2004, March). Curriculum-based measurement prediction of outcome on high stakes testing. Paper presented at the annual meeting of the National Association of School Psychologists, Dallas, TX.

Shapiro, E.S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum based measurement and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24*, 19 – 35.

Shinn, M.R., (2002). Best Practices in using curriculum-based measurement in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best Practices in school psychology IV* (pp. 671-698). Bethesda, MD: National Association of School Psychologists.

Shinn, M.R., (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford Press.

Shinn, M.R. (1995). Curriculum-based measurement and its use in a problem-solving model. In A. Thomas and J. Grimes (Eds.), *Best practices in school psychology III*. Washington DC: National Association of School Psychologists.

Shinn, M., & Shinn, M. (2002). AIMSweb Training Workbook: Administration and scoring of reading curriculum based measurement (R-CBM) for use in general outcome measurement. Edformation Inc. 1-42.

Schrank, F. A., Teglasi, H., Wolf, I. L., Miller, J. A., Caterino, L. C., Reynolds, C. R., et al. (2005). American Academy of School Psychology reply to response-to-intervention perspective. *The School Psychologist,* 59 (1), 30-33.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34,* 33-58, 78.

Underwood, B. J. (1961). Ten years of massed practice on distributed practice. *Psychological Review 4,* 229-247.

U. S. Department of Education (2004). Individuals with Disabilities Education Improvement Act of 2004. Retrieved on December 19, 2008, from

http://www.ed.gov/policy/speced/guid/idea/idea2004.html#law

U.S. Office of Education. (1968). *First annual report of the National Advisory Committee for Handicapped Children.* Washington, DC: U.S. Department of Health, Education, and Welfare.

U.S. Office of Education (1976, December 29). Proposed rule-making. *Federal Register*, *41*, (230), 52404–52407. Washington, DC: U.S. Government Printing Office.

U.S. Office of Education (1977, December 29). Assistance to states for education of handicapped children: Procedures for evaluating specific learning disabilities. *Federal Register, 42*, (250), 65082–65085. Washington, DC: U.S. Government Printing Office.

Woodcock, R.W. (1973). Woodcock reading mastery tests. Circle Pines, MN: American Guidance Service.

Zirkel, P.A. (2010). The legal meaning of specific learning disability for special education eligibility. *Teaching Exceptional Children, 42,* 62-67

Appendix

Appendix A: Recruitment Letter

Dear Parents,

Your child has been invited to participate in a research study that may help teachers better measure the progress of students learning how to read. Students who are in the second grade and who participate in the general education setting will be invited to participate in the study.

What would be asked of you?

- Agree to allow the researcher to measure your child's weekly reading progress for eight weeks.
- Agree to allow the researcher to share your child's progress with his/her teacher.
- Sign the consent form (attached) indicating interest in having your child participate.
- **That's all!**

This study is being conducted by Mrs. Dana Kelly, a doctoral student in psychology, as part of a dissertation effort for a doctoral degree in School Psychology from the University of Houston. She is being supervised by Dr. Thomas Kubiszyn, who can be reached at 713-743-9865. **Participation is strictly voluntary and would be very much appreciated!**
PLEASE RETURN TO YOUR CHILD'S TEACHER BY 3/5/2010

If you have questions, please call Dana Kelly at 713-743-9865.

This project has been reviewed by the University of Houston Committee for the Protection of Human Subjects (713) 743-9204.
This project is part of a dissertation effort sponsored by faculty member, Dr.Thomas Kubiszyn.

Appendix B: Parent Informed Consent

PLEASE SIGN LAST PAGE AND RETURN TO YOUR CHILD'S TEACHER

UNIVERSITY OF HOUSTON
PARENT PERMISSION TO PARTICIPATE IN RESEARCH

PROJECT TITLE:
The impact of data collection methods on reading fluency using second-grade Curriculum-Based Measurement-Reading (R-CBM) probes.

Your child is being invited to participate in a research project conducted by Dana Kelly. Ms. Kelly is a doctoral candidate from the Educational Psychology Department at the University of Houston. The project is part of a dissertation effort, and is conducted under the supervision of Dr. Thomas Kubiszyn.

NON-PARTICIPATION STATEMENT
Your child's participation is voluntary and you may refuse to participate or withdraw him/her at any time without penalty or loss of benefits to which you are otherwise entitled. You may also refuse to answer any question.

PURPOSE OF THE STUDY
The project is designed to help determine the effectiveness of different procedures used to collect data to measure reading fluency.

PROCEDURES
Potentially, approximately 200 children will participate in this study. All second grade children (whose parents consent) from the campuses will be assessed for reading fluency. Reading passages will be administered to each child individually by a trained individual. Each child's progress will be monitored during an eight-week period. Children will be removed from their classroom to be assessed. Each reading assessment will last approximately three to five minutes, so removal from instructional time will be minimal. At the end of the study, teachers and parents will receive information on each child's progress. Parent and teacher will be given a summary of the child's performance. In addition, the principle investigator will provide training to school staff regarding evidence-based practices in measuring reading fluency progress and in providing interventions. **All instruction and assessment will occur within the regular school day and are free of charge.**

ACCESS TO INFORMATION
Each parent will receive a summary of his/her child's performance, free of charge. If you provide permission on the last page of this form, teachers will be provided with the assessment results of your child's reading performance. The researcher will have access to each child's records in order to collect data and to assess progress.

CONFIDENTIALITY

Every effort will be made to maintain the confidentiality of your child's participation in this project. Your child's name will be paired with a code number. This code number will appear on all written materials. The list pairing your name to the assigned code number will be kept separate from all research materials and will be available only to the primary investigator and the reading specialist. Confidentiality will be maintained within legal limits.

RISKS/DISCOMFORTS
No risks or discomforts are associated with your participation in this study.

BENEFITS
Your child is expected to benefit from improved reading fluency skills as a result of participation in this study. The findings of the study may help investigators better understand how to better assess reading skills. This could lead to the development of more precise ways of measuring reading progress.

ALTERNATIVES
Participation in this project is voluntary and the only alternative to this project is non-participation.

PUBLICATION STATEMENT
The results of this study may be published in professional and/or scientific journals. It may also be used for educational purposes or for professional presentations. However, no individual subject will be identified.

SUBJECT RIGHTS

1. I understand that informed consent/permission is required for my child to participate in this project.
2. All procedures have been explained to me and all my questions have been answered to my satisfaction.
3. Any risks and/or discomforts for my child have been explained to me.
4. I understand that, if I have any questions, I may contact Dana Kelly at (713) 734-9864. I may also contact Dr. Thomas Kubiszyn, faculty sponsor, at (713) 743-9865.
5. I have been told that I may refuse for my child to participate or to stop my child's participation in this project at any time before or during the project. I may also refuse to answer any question.
6. ANY QUESTIONS REGARDING MY CHILD'S RIGHTS AS A RESEARCH SUBJECT MAY BE ADDRESSED TO THE UNIVERSITY OF HOUSTON COMMITTEE FOR THE PROTECTION OF HUMAN SUBJECTS (713-743-9204). ALL RESEARCH PROJECTS THAT ARE CARRIED OUT BY INVESTIGATORS AT THE UNIVERSITY OF HOUSTON ARE GOVERNED BY REQUIREMENTS OF THE UNIVERSITY AND THE FEDERAL GOVERNMENT.
7. All information that is obtained in connection with this project and that can be identified with my child will remain confidential as far as possible within legal limits. Information gained from this study that can be identified with my child may be released to no one other than the principal investigator and Dr. Thomas Kubiszyn. The results may be published in scientific journals, professional publications, or educational presentations without identifying my child by name.

I GIVE PERMISSION FOR MY CHILD'S ASSESSMENT RESULTS TO BE SHARED WITH THE TEACHER.

_____YES _____ NO

I HAVE READ (OR HAVE HAD READ TO ME) THE CONTENTS OF THIS PERMISSION FORM AND HAVE BEEN ENCOURAGED TO ASK QUESTIONS. I HAVE RECEIVED ANSWERS TO MY QUESTIONS. I GIVE MY PERMISSION FOR MY CHILD TO PARTICIPATE IN THIS STUDY. I HAVE RECEIVED (OR WILL RECEIVE) A COPY OF THIS FORM FOR MY RECORDS AND FUTURE REFERENCE.

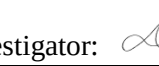
Study Subject (print child's name):_____

Signature of Parent:_____ Date: _____

I HAVE READ THIS FORM TO THE SUBJECT AND/OR THE SUBJECT HAS READ THIS FORM.

AN EXPLANATION OF THE RESEARCH WAS GIVEN AND QUESTIONS FROM THE SUBJECT WERE SOLICITED AND ANSWERED TO THE SUBJECT'S SATISFACTION. IN MY JUDGMENT, THE SUBJECT HAS DEMONSTRATED COMPREHENSION OF THE INFORMATION.

Principal Investigator: Dana E. Kelly

Signature of Principal Investigator:  Date: February 22, 2010

Appendix C: Teacher Consent To Participate

UNIVERSITY OF HOUSTON
TEACHER CONSENT TO PARTICIPATE IN RESEARCH

PROJECT TITLE:
The impact of data collection methods on reading fluency using second-grade Curriculum-Based Measurement-Reading (R-CBM) probes.

You are being invited to participate in a research project conducted by Dana Kelly. Ms. Kelly is a doctoral candidate from the Educational Psychology Department at the University of Houston. The project is part of a dissertation effort, and is conducted under the supervision of Dr. Thomas Kubiszyn.

NON-PARTICIPATION STATEMENT
Your participation is voluntary and you may refuse to participate or withdraw at any time without penalty or loss of benefits to which you are otherwise entitled. You may also refuse to answer any question.

PURPOSE OF THE STUDY
The project is designed to determine the measurement qualities of certain methods of measuring reading fluency.

PROCEDURES
Potentially, approximately 200 children will participate in this study. All second grade students (whose parents consent) from the campuses will be assessed for reading fluency. Reading passages will be administered to each student individually by a trained individual. Each student's progress will be monitored during an eight-week period. Students will be removed from their classroom to be assessed. Each reading assessment will last approximately three to five minutes, so removal from instructional time will be minimal. At the end of the study, teachers and parents will receive information on each student's progress. Parent and teacher will be given a summary of the student's performance. In addition, the principle investigator will provide training to school staff regarding evidence-based practices in measuring reading fluency progress and in providing interventions. **All instruction and assessment will occur within the regular school day and are free of charge.**

Approximately 10 teachers will be invited to participate in this project. Participants will meet with the principal investigator for a brief description of the study and an overview of assessment procedures.

After written parental consent is obtained, the reading specialist, who will collect the data will begin assessing each student individually. For eight weeks, the reading specialist will assess each student individually in a private setting. A daily schedule will be given to each teacher so that teachers will know which students will be assessed on any given day.

After conclusion of the study, participating teachers will be invited to attend a seminar about data collection for Curriculum-Based Measurement in Reading (R-CBM) for fluency and evidence-based interventions for reading fluency.

CONFIDENTIALITY
Every effort will be made to maintain the confidentiality of you and your students' participation in this project. Each student's name will be paired with a code number. This code number will appear on all written materials. The list pairing the student's name to the assigned code number will be kept separate from all research materials and will be available only to the primary investigator and the reading specialist. Confidentiality will be maintained within legal limits.

RISKS/DISCOMFORTS
No risks or discomforts are associated with your participation in this study.

BENEFITS
While you will not directly benefit from participation, your students are expected to benefit from improved reading fluency skills as a result of participation in this study. The findings of the study may help investigators better understand how to better assess reading skills. This could lead to the development of more precise ways of measuring reading progress. In addition, you will be given the opportunity to learn how to administer R-CBM fluency probes and interpret student performance. You will also be given some reading fluency interventions to use with your students that are supported by empirical research.

ALTERNATIVES
Participation in this project is voluntary and the only alternative to this project is non-participation.

FINANCIAL CONSIDERATION
Teachers who participate in the study will receive an American Express gift card in the amount of $25.00.

PUBLICATION STATEMENT
The results of this study may be published in professional and/or scientific journals. It may also be used for educational purposes or for professional presentations. However, no individual subject will be identified.

SUBJECT RIGHTS

1. I understand that informed consent is required of all persons participating in this project.

2. All procedures have been explained to me and all my questions have been answered to my satisfaction.

3. Any risks and/or discomforts have been explained to me.

4. Any benefits have been explained to me.

5. I understand that, if I have any questions, I may contact Dana Kelly at (713) 743-9865. I may also contact Dr. Thomas Kubiszyn, faculty sponsor, at (713) 743-9865.

6. I have been told that I may refuse to participate or to stop my participation in this project at any time before or during the project. I may also refuse to answer any question.

7. ANY QUESTIONS REGARDING MY RIGHTS AS A RESEARCH SUBJECT MAY BE ADDRESSED TO THE UNIVERSITY OF HOUSTON COMMITTEE FOR THE PROTECTION OF HUMAN SUBJECTS (713-743-9204). ALL RESEARCH PROJECTS THAT ARE CARRIED OUT BY INVESTIGATORS AT THE UNIVERSITY OF HOUSTON ARE GOVERNED BY REQUIREMENTS OF THE UNIVERSITY AND THE FEDERAL GOVERNMENT.

8. All information that is obtained in connection with this project and that can be identified with me will remain confidential as far as possible within legal limits. Information gained from this study that can be identified with me may be released to no one other than the principal investigator and Dr. Thomas Kubiszyn. The results may be published in scientific journals, professional publications, or educational presentations without identifying me by name.

I HAVE READ (OR HAVE HAD READ TO ME) THE CONTENTS OF THIS CONSENT FORM AND HAVE BEEN ENCOURAGED TO ASK QUESTIONS. I HAVE RECEIVED ANSWERS TO MY QUESTIONS. I GIVE MY CONSENT TO PARTICIPATE IN THIS STUDY. I HAVE RECEIVED (OR WILL RECEIVE) A COPY OF THIS FORM FOR MY RECORDS AND FUTURE REFERENCE.

Study Subject (print name):_____
Signature of Study Subject:_____ Date: _____

I HAVE READ THIS FORM TO THE SUBJECT AND/OR THE SUBJECT HAS READ THIS FORM. AN EXPLANATION OF THE RESEARCH WAS GIVEN AND QUESTIONS FROM THE SUBJECT WERE SOLICITED AND ANSWERED TO THE SUBJECT'S SATISFACTION. IN MY JUDGMENT, THE SUBJECT HAS DEMONSTRATED COMPREHENSION OF THE INFORMATION.

Principal Investigator: Dana E. Kelly

Signature of Principal Investigator:  Date: February 22, 2010

Appendix D: Student Assent

Dear Second Grade Student,

I am a person who works with children to help them learn better. I am going to school, too, to learn more about helping children in school.

For my homework, I am trying to find out more about how second-grade students read. Your mom or dad has said it's ok for you to come and read with me or one of my helpers during school.

If you would like to read to me or one of my helpers for 6 weeks, please circle "Yes" below. If you don't want to do this, circle "No." Either way is ok.

YES NO

Print your name here_____

Dana Kelly _____
Researcher's name Date

Thank You!

Appendix E: Outline for Principal/Staff Meeting

Introductions

PROJECT TITLE:
The impact of data collection methods on reading fluency using second-grade Curriculum-Based Measurement-Reading (CBM-R) probes.

NON-PARTICIPATION STATEMENT
Your participation is voluntary and you may refuse to participate or withdraw at any time without penalty or loss of benefits to which you are otherwise entitled. You may also refuse to answer any question.

PURPOSE OF THE STUDY
The project is designed to determine the measurement qualities of certain methods of measuring reading fluency.

PROCEDURES
All second grade students (whose parents consent) from the campuses will be assessed for reading fluency. Reading passages will be administered to each child individually by a trained reading specialist. Each student's progress will be monitored during a six-week period. Students will be removed from their classroom to be assessed by a data collector. Each reading assessment will last approximately one minute, so removal from instructional time will be minimal. At the end of the study, teachers and parents will receive information on each child's progress. Parent and teacher will be given a summary of the child's performance. In addition, the principle investigator will provide training to school staff regarding evidence-based practices in measuring reading fluency progress and in providing interventions. **All instruction and assessment will occur within the regular school day and are free of charge.**

Approximately 10 teachers will be invited to participate in this project. Participants will meet with the principal investigator for a brief description of the study and an overview of assessment procedures.

After written parental consent is obtained, the data collectors will begin assessing each student individually. For six weeks, the data collectors will assess each student individually in a private setting. A daily schedule will be given to each teacher so that disruption to class will be minimal.

After conclusion of the study, participating teachers will be invited to attend a seminar about data collection for Curriculum-Based Measurement in Reading (CBM-R) for fluency and evidence-based interventions for reading fluency.

CONFIDENTIALITY

Every effort will be made to maintain the confidentiality of you and your students' participation in this project. Each student's name will be paired with a code number. This code number will appear on all written materials. The list pairing the student's name to the assigned code number will be kept separate from all research materials and will be available only to the primary investigator and the reading specialist. Confidentiality will be maintained within legal limits.

RISKS/DISCOMFORTS

No risks or discomforts are associated with your participation in this study. You may have some students who may experience some anxieties regarding taking tests, however, they will be administered reading probes, which is what has already occurred in your class.

BENEFITS

While you will not directly benefit from participation, your students are expected to benefit from improved reading fluency skills as a result of participation in this study. The findings of the study may help investigators better understand how to better assess reading skills. This could lead to the development of more precise ways of measuring reading progress. In addition, you will be given the opportunity to learn how to administer CBM-R fluency probes and interpret student performance. You will also be given some reading fluency interventions to use with your students that are supported by empirical research.

ALTERNATIVES

Participation in this project is voluntary and the only alternative to this project is non-participation.

FINANCIAL CONSIDERATION

Teachers who participate in the study will receive a gift card in the amount of $25.00.

PUBLICATION STATEMENT

The results of this study may be published in professional and/or scientific journals. It may also be used for educational purposes or for professional presentations. However, no individual subject will be identified.

Appendix F: Reading-Curriculum-Based Measurement Probes

DIBELS Probe Example

ORF Progress Monitoring 1

Riding the Bus to School

I ride a big yellow bus to school. I stand on the corner of our street with my friends and we wait for the bus. My friend's grandma waits with us. When it's raining, she holds an umbrella to keep us dry. Sometimes when it's cold she brings us hot chocolate.

I leave my house to walk to the bus stop after my parents go to work. I watch the clock so I know when to leave. Sometimes mom phones me from her office to remind me. Sometimes she can't call, so I have to be sure to watch the time.

Our bus driver puts his flashing yellow lights on and then stops right next to us. When he has stopped he turns the red lights on so all the cars will stop. He makes sure we are all sitting down before he starts to go. He watches out for us very carefully.

My friends and I are the first ones to be picked up by the bus. We like to sit right behind the bus driver and watch while he picks up all the other kids.We know where everyone lives. By the time we get to our school, the bus is almost full. Sometimes the kids get noisy and the driver has to remind us to keep it

down. He says their noise makes it hard for him to concentrate and drive safely. I am glad that our bus driver is so careful.

AIMSWeb Probe Example

At my house, Friday night is family night. Our whole family gets together to do something fun. Two weeks ago we went bowling. Last Friday we went to an art show. This week we planned to see a movie at the movie theater.

"What movie shall we see?" Dad asked.

"I like action movies," my brother said. "I like to watch cars crash. I like to watch super-heroes fly."

"I like animal movies," my sister said. "I want to see horses run free in fields. I want to see whales swim in the sea."

"I like funny movies," Dad said. "I laugh when people throw pies. I laugh when people tell funny jokes."

"I like movies about love," Mom said. "I like it when a man and a woman get married and live happily ever after."

"I like cartoons," I said. "I like colorful movies with a lot of music."

What could we do? Our family could not choose a movie to watch together.

Dad thought he'd solve the problem. He said, "Why don't we stay home and play a family game?" We all thought that was a good idea.

"Let's play puzzles!" I said.

"Let's play cards!" my brother said.

"Let's play checkers!" my sister said.

Dad just shook his head and rolled his eyes. "I'll be in bed," he said. "Wake me when family night begins."

Appendix G

Tables

Table 1

Condition A Median

Student	Week 1	Week 2	Week 3	Week 4	Week 5	Week 1
1	99	92	99	102	106	99
2	74	78	76	81	70	74
3	143	146	129	126	130	143
4	154	150	152	166	160	154
5	155	138	132	127	132	155
6	126	135	121	141	123	126
7	125	110	118	116	117	125
8	140	148	133	152	140	140
9	99	98	96	98	99	99
10	74	78	77	77	77	74
11	68	64	75	69	59	68
12	107	108	95	105	100	107
13	110.5	101	94	92	105	110.5
14	96	74	93	105	89	96
15	132	141	136	139	145	132
16	112	99	95	105	95	112
17	117	105	101	111	109	117
18	85	77	73	74	74	85
19	126	128	125	112	120	126
20	136	125	131	128	133	136
21	116	107	92	103	93	116
22	137	136	138	130	122	137
23	91.5	81	90	85	87	91.5
24	97	101	103	91	97	97
25	87	80	95	98	77	87
26	104	105	102	100	91	104
27	100	96	99	99	92	100
28	99	103	110	100	95	99
29	138	132	130	131	116	138
30	82	79	78	80	67	82
31	73	61	73	76	76	73
32	102	99	106	97	108	102
33	132	126	136	128	119	132
34	99	89	90	86	92	99
35	123	119	138	123	118	123
36	81	83	85	89	86	81
37	92	100	100	103	89	92
38	121	135	125	138	124	121
39	73	58	67	73	68	73
40	94	86	85	88	88	94
41	130	124	129	128	126	130
42	133	132	126	128	132	133

Table 2

Condition B Median

Student	Week 1	Week 2	Week 3	Week 4	Week 5	Week 1
1	99	95	99	102	105	99
2	83	85	72	92	83	96
3	152	146	133	126	128	133
4	154	153	157	163	158	151
5	160	151	141	121	131	134
6	127	137	131	139	126	117
7	131	110	117	116	131	105
8	132	158	139	158	141	150
9	99	101	110	97	106	100
10	83	80	81	66	87	71
11	68	62	75	70	68	74
12	101	108	99	102	92	99
13	117	103	96	102	110	101
14	103	65	97	110	101	79
15	135	143	128	135	145	136
16	126	92	94	110	105	96
17	106	111	90	111	114	108
18	85	77	73	74	76	77
19	134	129	130	112	117	122
20	122	129	127	128	125	121.5
21	116	94	74	103	97	105
22	137	129	147	130	136	145
23	103	95	112	77	96	86
24	97	103	109	99	97	89
25	96	79	89	99	82	78
26	106	106	108	105	93	104
27	107	106	99	104	101	93
28	100	109	104	99	100	74
29	123	117	130	123	127	125
30	79	78	87	83	79	74
31	73	74	73	74	79	72
32	98	106	114	95	111	111
33	132	134	137	138	115	113
34	99	97	111	92	87	107
35	119	132	145	117	118	130
36	105	95	76	89	96	92
37	91	101	104	98	85	89
38	121	128	145	129	135	139
39	77	78	60	64	78	71
40	92	103	91	88	88	100
41	127	129	131	133	126	134
42	136	136	129	128	132	131.5

Table 3

Condition C Median

Student	Week 1	Week 2	Week 3	Week 4	Week 5	Week 1
1	99	104	87	85	106	102
2	70	76	81	73	85	96
3	143	153	129	126	131	161
4	151	150	169	166	160	161
5	144	155	132	137	131	134
6	115	145	124	142	128	127
7	103	116	118	109	131	116
8	133	148	138	163	141	155
9	98	99	97	98	97	109
10	67	83	81	66	86	87
11	68	60	75	65	58	62
12	107	116	93	115	92	113
13	108	107	90	91	112	119
14	71	101	97	110	101	97
15	137	145	152	139	157	142
16	112	91	94	105	86	112
17	117	106	101	102	110	123
18	88	77	79	74	73	77
19	126	135	119	102	120	144
20	134	144	135	135	133	148
21	104	118	109	92	93	115
22	137	136	123	135	136	153
23	91	89	103	82	103	99
24	97	104	105	91	108	111
25	81	112	89	85	74	98
26	98	113	101	100	88	103
27	89	103	98	82	88	101
28	89	113	114	97	105	111
29	141	133	140	123	116	133
30	78	79	87	80	71	85
31	67	77	66	74	76	80
32	102	101	107	97	113	108
33	132	126	136	117	115	139
34	109	107	88	83	85	113
35	134	136	138	123	110	137
36	74	83	98	84	86	103
37	91	107	106	91	91	95
38	118	154	132	129	116	139
39	57	57	69	64	63	85
40	88	104	79	83	93	105
41	130	135	141	126	128	126
42	138	136	140	128	131	139

Table 4

Condition D Median

Student	Week 1	Week 2	Week 3	Week 4	Week 5	Week 1
1	99	104	106	85	105	99
2	75	85	81	74	88	96
3	130	153	129	126	146	161
4	154	150	192	166	150	161
5	144	161	154	121	131	134
6	115	145	131	138	131	127
7	103	135	121	103	137	105
8	132	167	158	163	141	159
9	98	112	110	97	112	104
10	84	83	81	66	87	85
11	68	60	75	70	58	81
12	100	129	93	115	92	107
13	107	120	96	91	127	119
14	95	116	97	110	101	76
15	137	145	155	135	157	149
16	112	91	94	105	120	112
17	106	111	108	101	122	123
18	88	77	79	74	73	77
19	115	143	138	102	117	151
20	103	144	136	144	125	131
21	88	108	109	91	97	89
22	109	138	147	125	136	150
23	101	114	119	64	127	99
24	97	104	109	91	111	111
25	96	112	89	98	74	98
26	98	116	116	111	104	104
27	89	106	103	82	101	101
28	80	121	125	99	105	111
29	123	154	144	123	138	133
30	78	82	87	80	79	85
31	72	77	65	74	76	80
32	98	106	129	95	113	102
33	132	143	136	121	115	139
34	109	116	111	83	85	113
35	119	136	145	99	126	141
36	81	95	99	89	86	108
37	91	107	106	97	96	87
38	118	154	145	129	124	139
39	65	88	60	64	78	77
40	88	105	100	83	93	105
41	115	142	143	119	128	138
42	129	136	147	128	131	130

Table 5

Condition A Mean

Student	Week 1	Week 2	Week 3	Week 4	Week 5	Week 1
1	101.44	91.22	95.89	99.22	98.78	98.33
2	74.11	79.56	74.56	80.89	73.00	88.44
3	140.33	148.00	130.89	125.89	128.89	143.00
4	152.44	146.89	153.78	167.78	158.00	156.78
5	160.67	142.33	133.00	127.67	131.33	133.56
6	125.78	133.67	120.89	141.00	118.56	121.89
7	120.00	112.67	114.22	116.56	114.67	112.33
8	141.67	138.67	132.00	151.56	135.44	145.78
9	98.33	96.67	97.00	99.78	98.33	100.78
10	73.89	77.22	75.00	74.00	76.22	77.22
11	68.33	63.78	69.67	67.22	57.78	66.56
12	109.44	105.44	99.44	104.56	101.44	102.33
13	109.00	97.89	97.89	96.78	104.22	102.44
14	93.56	83.33	95.00	107.78	93.78	93.44
15	129.89	141.78	139.00	138.22	141.11	142.00
16	112.33	97.44	92.00	107.22	95.22	97.33
17	113.33	103.89	95.22	114.33	107.67	106.22
18	80.67	78.56	73.89	75.78	74.56	79.11
19	127.22	127.56	125.00	115.44	122.33	129.00
20	135.33	129.33	126.22	130.78	135.44	136.17
21	108.11	102.44	89.33	103.67	88.89	121.00
22	133.56	131.67	137.11	131.78	125.56	142.11
23	89.83	83.56	90.67	84.78	88.33	89.78
24	99.78	100.89	104.56	93.11	98.00	97.78
25	89.33	86.78	87.11	97.44	76.89	88.56
26	103.22	100.67	103.33	98.56	92.00	96.89
27	95.33	95.44	97.89	96.33	92.11	95.44
28	95.33	105.22	104.22	102.00	96.56	95.67
29	132.00	129.22	128.67	131.22	105.56	128.11
30	82.11	76.67	78.44	163.88	67.33	77.78
31	70.78	66.11	72.22	79.00	71.00	74.44
32	101.67	96.89	109.56	102.11	101.67	108.11
33	133.89	122.00	133.78	127.56	116.33	122.33
34	98.11	92.89	98.33	93.67	92.89	99.00
35	125.22	121.33	133.56	121.67	119.00	124.78
36	85.44	82.22	88.67	91.56	81.11	96.00
37	90.44	100.89	98.22	99.44	88.11	95.22
38	123.33	134.00	126.56	133.44	122.67	139.67
39	71.22	62.11	69.00	73.67	68.33	72.56
40	94.00	87.56	85.67	87.22	88.22	95.56
41	130.11	125.00	130.11	128.56	122.33	127.00
42	132.22	131.56	125.22	133.33	131.56	138.40

Table 6

Condition B Mean

Student	Week 1	Week 2	Week 3	Week 4	Week 5	Week 1
1	101.67	94.67	91.67	97.00	100.33	92.33
2	84.33	84.00	69.33	86.67	80.33	88.00
3	147.67	142.67	135.33	119.00	133.67	141.67
4	155.33	153.00	166.00	154.00	157.67	154.00
5	162.67	153.67	140.33	118.00	127.67	134.67
6	128.00	139.00	123.33	140.00	127.00	117.67
7	122.00	116.33	107.00	118.00	118.67	107.33
8	128.67	157.67	128.00	156.00	133.33	140.33
9	102.00	103.33	103.00	100.33	101.33	95.33
10	83.33	77.67	76.00	70.33	82.67	73.00
11	71.67	64.33	66.33	73.67	65.00	69.67
12	109.33	113.00	107.00	101.67	105.00	100.33
13	117.00	102.00	102.67	98.33	111.00	101.00
14	103.67	76.67	97.67	108.00	97.00	83.00
15	134.67	143.00	132.33	132.67	143.33	138.67
16	124.00	94.00	88.33	108.33	108.33	95.67
17	113.33	109.67	92.00	113.67	113.67	106.33
18	85.00	78.33	70.67	71.00	78.67	75.67
19	131.67	133.33	127.00	115.67	118.00	126.33
20	121.67	127.67	120.33	133.00	133.67	121.50
21	107.33	97.33	85.33	103.33	94.00	105.00
22	130.67	125.00	145.67	128.33	126.67	143.67
23	103.00	93.00	99.00	80.33	100.00	84.00
24	101.33	101.33	110.00	99.33	101.67	91.33
25	94.67	86.67	80.67	102.67	79.67	84.33
26	105.33	109.00	104.67	105.00	95.33	104.00
27	102.33	104.00	94.33	98.67	95.67	94.67
28	98.33	111.00	100.33	98.00	98.33	85.67
29	124.67	128.33	126.67	125.00	114.33	121.00
30	83.67	74.67	81.67	85.00	80.33	75.67
31	74.67	70.67	73.00	80.67	79.00	70.00
32	103.00	98.67	115.67	104.33	110.67	109.00
33	134.00	128.67	138.67	133.67	114.00	114.67
34	102.33	98.33	112.33	98.33	94.00	99.33
35	121.00	125.33	136.67	116.00	117.33	127.33
36	99.00	92.67	83.00	90.67	93.33	93.00
37	92.67	102.00	93.33	99.33	87.67	93.33
38	123.67	135.67	132.67	130.33	134.67	133.00
39	77.33	71.00	69.00	68.33	73.33	66.00
40	95.00	98.33	92.00	91.33	88.33	98.00
41	126.00	128.00	130.00	129.33	120.00	130.33
42	134.67	132.33	126.67	126.00	132.33	131.50

Table 7

Condition C Mean

Student	Week 1	Week 2	Week 3	Week 4	Week 5	Week 1
1	103.33	102.00	92.67	84.33	106.67	101.67
2	71.67	75.33	82.33	70.67	79.00	95.00
3	144.33	152.67	126.00	131.00	129.00	161.33
4	150.00	153.67	175.33	170.00	159.00	167.33
5	151.00	150.33	138.00	135.00	130.67	135.33
6	121.67	132.33	123.33	146.67	127.33	127.67
7	109.33	118.67	118.67	108.00	123.00	120.33
8	140.33	150.00	141.67	159.33	137.67	151.67
9	96.00	95.00	101.00	102.33	99.33	108.33
10	67.67	79.00	78.67	69.33	81.00	88.00
11	65.33	63.67	76.33	66.67	53.67	64.00
12	111.67	115.00	91.67	114.67	90.67	112.33
13	109.33	107.33	90.67	88.67	112.67	115.67
14	78.33	97.00	96.00	111.67	102.00	93.33
15	137.67	141.00	143.00	139.00	151.33	143.00
16	105.67	91.00	92.67	108.33	94.67	104.67
17	113.33	103.33	103.33	103.33	103.00	121.00
18	80.00	77.67	77.00	77.00	69.67	79.33
19	131.00	133.00	124.00	99.33	125.67	141.00
20	127.33	137.67	134.00	132.33	131.67	144.00
21	110.33	115.00	107.67	93.67	87.00	108.67
22	138.67	132.67	127.33	133.00	132.00	152.00
23	87.33	97.00	101.00	77.00	103.67	99.33
24	95.33	106.00	102.33	87.67	104.33	113.67
25	86.00	108.00	89.00	88.00	70.33	99.00
26	104.33	113.67	102.67	102.00	89.67	101.00
27	91.67	97.00	99.67	86.00	85.33	100.00
28	89.00	111.33	114.00	96.33	104.33	110.00
29	136.00	136.33	133.00	127.00	105.33	136.33
30	80.67	78.00	86.67	301.33	68.00	84.00
31	64.67	72.00	68.33	73.67	71.33	84.33
32	103.00	95.33	114.00	98.00	103.00	108.67
33	137.00	129.33	131.67	110.33	115.67	134.00
34	102.67	102.67	95.00	81.00	87.67	116.00
35	134.33	127.00	133.33	116.67	115.00	136.67
36	68.67	77.00	98.00	83.00	79.67	103.00
37	85.33	106.33	104.67	92.00	82.00	93.67
38	121.33	148.67	129.33	118.00	114.67	147.67
39	56.33	66.33	71.67	65.00	66.67	82.33
40	92.00	98.33	82.67	81.67	92.00	104.00
41	129.67	132.00	137.00	124.33	123.33	129.67
42	135.33	134.67	137.67	135.33	132.00	138.00

Table 8

Condition D Mean

Student	Week 1	Week 2	Week 3	Week 4	Week 5	Week 1
1	99	104	106	85	105	99
2	75	85	81	74	88	96
3	130	153	129	126	146	161
4	154	150	192	166	150	161
5	144	161	154	121	131	134
6	115	145	131	138	131	127
7	103	135	121	103	137	105
8	132	167	158	163	141	159
9	98	112	110	97	112	104
10	84	83	81	66	87	85
11	68	60	75	70	58	81
12	100	129	93	115	92	107
13	107	120	96	91	127	119
14	95	116	97	110	101	76
15	137	145	155	135	157	149
16	112	91	94	105	120	112
17	106	111	108	101	122	123
18	88	77	79	74	73	77
19	115	143	138	102	117	151
20	103	144	136	144	125	131
21	88	108	109	91	97	89
22	109	138	147	125	136	150
23	101	114	119	64	127	99
24	97	104	109	91	111	111
25	96	112	89	98	74	98
26	98	116	116	111	104	104
27	89	106	103	82	101	101
28	80	121	125	99	105	111
29	123	154	144	123	138	133
30	78	82	87	80	79	85
31	72	77	65	74	76	80
32	98	106	129	95	113	102
33	132	143	136	121	115	139
34	109	116	111	83	85	113
35	119	136	145	99	126	141
36	81	95	99	89	86	108
37	91	107	106	97	96	87
38	118	154	145	129	124	139
39	65	88	60	64	78	77
40	88	105	100	83	93	105
41	115	142	143	119	128	138
42	129	136	147	128	131	130

Table 11

Tests of Normality

	Kolomoforov-Smirnov (a)			Shapiro-Wilk		
	Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
W1C1	.102	42	.200(*)	.966	42	.237
W2C1	.101	42	.200(*)	.962	42	.172
W3C1	.135	42	.054	.946	42	.046
W4C1	.132	42	.064	.966	42	.232
W5C1	.092	42	.200(*)	.980	42	.673
W6C1	.150	42	.018	.967	42	.254
W1C2	.114	42	.200(*)	.977	42	.548
W2C2	.114	42	.192	.970	42	.328
W3C2	.125	42	.097	.967	42	.255
W4C2	.085	42	.200(*)	.981	42	.683
W5C2	.100	42	.200(*)	.972	42	.391
W6C2	.088	42	.200(*)	.951	42	.071
W1C3	.114	42	.196	.959	42	.131
W2C3	.112	42	.200(*)	.961	42	.161
W3C3	.097	42	.200(*)	.966	42	.246
W4C3	.106	42	.200(*)	.956	42	.107
W5C3	.099	42	.200(*)	.980	42	.660
W6C3	.094	42	.200(*)	.979	42	.632
W1C4	.078	42	.200(*)	.984	42	.821
W2C4	.112	42	.200(*)	.971	42	.365
W3C4	.092	42	.200(*)	.982	42	.736
W4C4	.097	42	.200(*)	.964	42	.203
W5C4	.094	42	.200(*)	.979	42	.636
W6C4	.115	42	.191	.952	42	.074

* This is a lower bound of the true significance.

a Lilliefors Significance Correction

Table 19

Weekly Median Scores (of all 9 Probes), Change over Six Weeks, and Median of Medians

Student	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Total Change	AIMSweb Change	DIBELS Change
1	99	92	99	102	106	102	3	0	0
2	74	78	76	81	70	92	18	2	11
3	143	146	129	126	130	140	-3	-14	14
4	154	150	152	166	160	161	7	-2	-5
5	155	138	132	127	132	134	-21	-23	7
6	126	135	121	141	123	124	-2	-5	-17
7	125	110	118	116	117	114	-11	-7	-2
8	140	148	133	152	140	150	10	-7	-2
9	99	98	96	98	99	100	1	-3	2
10	74	78	77	77	77	85	11	3	8
11	68	64	75	69	59	62	-6	7	-7
12	107	108	95	105	100	100	-7	-12	-5
13	110.5	101	94	92	105	99	-11.5	-16.5	7
14	96	74	93	105	89	94	-2	-3	-11
15	132	141	136	139	145	142	10	4	3
16	112	99	95	105	95	96	-16	-17	-9
17	117	105	101	111	109	103	-14	-16	-8
18	85	77	73	74	74	77	-8	-12	3
19	126	128	125	112	120	128	2	-1	16
20	136	125	131	128	133	136.5	0.5	-5	8.5
21	116	107	92	103	93	121.5	5.5	-24	18.5
22	137	136	138	130	122	145	8	1	15
23	91.5	81	90	85	87	86	-5.5	-1.5	1
24	97	101	103	91	97	102	5	6	11
25	87	80	95	98	77	82	-5	8	-16
26	104	105	102	100	91	99	-5	-2	-1
27	100	96	99	99	92	93	-7	-1	-6
28	99	103	110	100	95	101	2	11	1
29	138	132	130	131	116	127	-11	-8	-4
30	82	79	78	80	67	78	-4	-4	-2
31	73	61	73	76	76	80	7	0	4
32	102	99	106	97	108	109	7	4	12
33	132	126	136	128	119	120	-12	4	-8
34	99	89	90	86	92	107	8	-9	21
35	123	119	138	123	118	125	2	15	2
36	81	83	85	89	86	98	17	4	9
37	92	100	100	103	89	95	3	8	-8
38	121	135	125	138	124	139	18	4	1
39	73	58	67	73	68	77	4	-6	4
40	94	86	85	88	88	94	0	-9	6
41	130	124	129	128	126	126	-4	-1	-2
42	133	132	126	128	132	139	6	-7	11
Median of Medians	**103**	**102**	**100.5**	**103**	**100.5**	**102.5**			

Table 20

Range of Performance for Each Probe

Probe	Low	High	Range	Probe	Low	High	Range
1.1.1	79	184	105	4.1.1	58	158	100
1.1.2	66	155	89	4.1.2	61	179	118
1.1.3	63	193	130	4.1.3	46	155	109
1.2.1	65	154	89	4.2.1	64	166	102
1.2.2	47	145	98	4.2.2	60	185	125
1.2.3	57	173	116	4.2.3	65	159	94
1.3.1	67	166	99	4.3.1	69	163	94
1.3.2	64	154	90	4.3.2	59	185	126
1.3.3	52	158	106	4.3.3	67	185	118
2.1.1	47	156	109	5.1.1	61	165	104
2.1.2	48	149	101	5.1.2	64	162	98
2.1.3	56	148	92	5.1.3	48	174	126
2.2.1	60	167	107	5.2.1	58	157	99
2.2.2	57	172	115	5.2.2	44	167	123
2.2.3	51	143	92	5.2.3	56	160	104
2.3.1	49	158	109	5.3.1	68	158	90
2.3.2	58	153	95	5.3.2	53	158	105
2.3.3	64	170	106	5.3.3	48	129	81
3.1.1	45	149	104	6.1.1	67	150	83
3.1.2	45	152	107	6.1.2	72	170	98
3.1.3	75	162	87	6.1.3	81	175	94
3.2.1	60	192	132	6.2.1	76	161	85
3.2.2	71	169	98	6.2.2	62	180	118
3.2.3	64	165	101	6.2.3	49	161	112
3.3.1	72	157	85	6.3.1	50	151	101
3.3.2	58	153	95	6.3.2	57	135	78
3.3.3	58	144	86	6.3.3	56	151	95

Note. Probe column is in the format: Week.day.probe#

Table 21

Relationship Between Readability, Median Performance, and Mean of Each Probe

Probe	Spache	Median	Mean	Probe	Spache	Median	Mean
1.1.1	2.3	112.0	112.9	4.1.1	2.7	102.0	103.6
1.1.2	2.2	98.0	119.2	4.1.2	2.7	104.0	108.3
1.1.3	2.4	114.0	115.2	4.1.3	3.0	100.0	101.7
1.2.1	2.7	99.5	102.8	4.2.1	2.9	99.0	102.5
1.2.2	2.5	97.5	98.3	4.2.2	3.2	95.0	103.2
1.2.3	2.5	114.5	115.8	4.2.3	2.4	100.0	103.9
1.3.1	2.4	114.0	115.4	4.3.1	2.8	113.0	115.3
1.3.2	2.1	104.0	107.7	4.3.2	2.7	108.0	112.0
1.3.3	2.2	96.0	100.0	4.3.3	2.7	118.0	117.2
2.1.1	2.7	103.5	102.4	5.1.1	2.7	93.0	98.2
2.1.2	2.3	99.5	99.2	5.1.2	2.5	106.0	108.6
2.1.3	3.4	94.0	96.3	5.1.3	2.7	101.0	103.4
2.2.1	2.1	116.0	118.0	5.2.1	2.6	114.0	111.0
2.2.2	2.7	108.0	113.0	5.2.2	2.4	107.5	106.5
2.2.3	2.2	102.0	99.8	5.2.3	2.7	88.5	92.7
2.3.1	2.6	103.0	104.3	5.3.1	2.7	108.0	109.8
2.3.2	2.3	100.5	101.7	5.3.2	3.2	100.5	101.0
2.3.3	2.1	107.5	109.7	5.3.3	3.0	91.5	88.0
3.1.1	2.9	90.0	90.5	6.1.1	2.8	102.0	102.8
3.1.2	2.6	99.5	97.6	6.1.2	2.8	111.0	114.5
3.1.3	3.0	111.5	114.4	6.1.3	2.8	119.0	121.8
3.2.1	2.2	110.5	115.7	6.2.1	2.9	111.0	113.9
3.2.2	1.8	102.0	104.4	6.2.2	2.7	115.5	118.2
3.2.3	2.4	101.0	106.5	6.2.3	2.5	109.0	113.4
3.3.1	2.2	114.0	113.1	6.3.1	3.2	92.0	96.8
3.3.2	2.6	105.5	105.1	6.3.2	2.7	90.0	92.1
3.3.3	2.4	103.5	101.2	6.3.3	3.1	89.0	91.8

Note. Probe column is in the format: Week.day.probe#

Table 22

Relationship between DRA level and Universal Screening/Survey-level assessment Median Score

Student	DRA	Median	Student	DRA	Median
1	30	92	22	30	132
2	28	83	23	20	92
3	40	147	24	24	104
4	40	155	25	28	101
5	40	184	26	24	112
6	40	125	27	38	105
7	40	121	28	24	100
8	40	120	29	38	138
9	38	99	30	24	85
10	24	72	31	30	78
11	24	74	32	34	103
12	30	101	33	34	121
13	34	108	34	30	99
14	20	107	35	34	114
15	34	133	36	24	97
16	30	123	37	24	100
17	38	120	38	28	132
18	28	82	39	20	88
19	24	130	40	24	96
20	38	136	41	16	127
21	30	117	42	30	127

Table 23

Weekly Median Scores of Students Aggregated by Teacher

Teacher	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Total Change	AIMSweb Change	DIBELS Change
A	107	108	95	105	100	100	-7	-12	-5
A	116	107	92	103	93	121.5	5.5	-24	18.5
A	137	136	138	130	122	145	8	1	15
A	102	109	120	110	116	127	25	18	17
A	81	83	85	89	86	98	17	4	9
A	133	132	126	128	132	139	6	-7	11
B	143	146	129	126	130	140	-3	-14	14
B	154	150	152	166	160	161	7	-2	-5
B	155	138	132	127	132	134	-21	-23	7
B	126	135	121	141	123	124	-2	-5	-17
B	125	110	118	116	117	114	-11	-7	-2
B	140	148	133	152	140	150	10	-7	-2
B	99	98	96	98	99	100	1	-3	2
B	126	128	125	112	120	128	2	-1	16
B	136	125	131	128	133	136.5	0.5	-5	8.5
C	74	78	76	81	70	92	18	2	11
C	74	78	77	77	77	85	11	3	8
C	68	64	75	69	59	62	-6	7	-7
C	99	99	96		103.5	107	8	-3	-3.5
C	97	101	103	91	97	102	5	6	11
C	87	80	95	98	77	82	-5	8	-16
C	104	105	102	100	91	99	-5	-2	-1
C	100	96	99	99	92	93	-7	-1	-6
C	99	103	110	100	95	101	2	11	1
C	138	132	130	131	116	127	-11	-8	-4
C	82	79	78	80	67	78	-4	-4	-2
C	92	100	100	103	89	95	3	8	-8
C	113	112	110	115	101	104	-9	-3	-11
C	121	135	125	138	124	139	18	4	1
D	99	92	99	102	106	102	3	0	0
D	110.5	101	94	92	105	99	-11.5	-16.5	7
D	96	74	93	105	89	94	-2	-3	-11
D	84	70	80	75	73	74	-10	-4	-1
D	91.5	81	90	85	87	86	-5.5	-1.5	1
D	73	58	67	73	68	77	4	-6	4
D	94	86	85	88	88	94	0	-9	6
E	132	141	136	139	145	142	10	4	3
E	112	99	95	105	95	96	-16	-17	-9
E	117	105	101	111	109	103	-14	-16	-8
E	85	77	73	74	74	77	-8	-12	3
F	73	61	73	76	76	80	7	0	4
F	102	99	106	97	108	109	7	4	12
F	132	126	136	128	119	120	-12	4	-8
F	99	89	90	86	92	107	8	-9	21
F	123	119	138	123	118	125	2	15	2
F	130	124	129	128	126	126	-4	-1	-2