To the Graduate Council:

I am submitting herewith a dissertation written by Brian Charles Poncy entitled "An Investigation of the Dependability and Standard Error of Measurement of Words Read Correctly Per Minute Using Curriculum-Based Measurement." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Education.

Christopher H. Skinner
Major Professor

We have read this dissertation
and recommend its acceptance:

R. Steve McCallum

Schuyler W. Huck

Richard A. Saudargas

Accepted for the Council:

Anne Mayhew
Vice Chancellor and
Dean of Graduate Studies

(Original signatures are on file with official student records)

AN INVESTIGATION OF THE DEPENDABILITY AND STANDARD ERROR OF MEASUREMENT OF WORDS READ CORRECTLY PER MINUTE USING CURRICULUM-BASED MEASUREMENT

A dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Brian Charles Poncy
August 2006

ABSTRACT

Generalizability (G) theory was used in two studies to assess the variability in words correct per minute (wcpm) scores caused by student skill and passage variability. Study one was a small "n" study with a sample of 14 third-grade students and study 2 had a sample of 37 third-grade students. Reliability-like coefficients and the SEM based on a specific number of assessments using different combinations of passages demonstrated how manipulating probe variability could reduce measurement error. Results of the two studies showed that 69% and 81% of the variance was due to student skill, 20% and 10% of the variance was due to passage or probe variability, and 10% and 9% of the variance was due to unaccounted sources of error. Reliability-like coefficients ranged from .68 to .99 and SEMs ranged from 18 to 4 wcpm depending on the number of probes given. When passage variability was controlled, SEMs were decreased and ranged from 12 to 4 wcpm. Results indicated that wcpm scores yield high reliability-like coefficients, but also have a large SEM that can be reduced by administering multiple alternate passages. Discussion focuses on conducting research designed to identify more equivalent passages in order to reduce erroneous relative and absolute decisions.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

LITERATURE REVIEW

The reauthorization of IDEA has prompted discussion concerning the role and function of school psychologists and the need to implement an often discussed paradigm shift (Reschly & Ysseldyke, 2002; Ysseldyke & Marston, 1998). A topic that continues to be debated is whether practitioners should use traditional assessment approaches within a discrepancy model or use functional assessments in a problem solving model to determine special education eligibility (Gresham & Noell, 1998; Gresham & Witt, 1997; Shinn, Good, & Parker, 1998). In response to this debate, the National Association of School Psychologists (NASP) (2003) has recommended the cessation of the ability-achievement discrepancy requirement for entitlement purposes. As an alternative method to special education entitlement, NASP (2003) has suggested that a variety of assessments be used in a problem solving model to determine if students display significantly low achievement and insufficient response to intervention. An efficient method to gather information concerning a student's level of academic achievement and his/her responsiveness to intervention is through the use of Curriculum-Based Measurement (CBM), a set of standardized procedures used to measure basic academic skills (Shinn, 1989).

The Paradigm Shift: From ATI to Problem Solving

Traditionally, norm referenced assessments have been used in a pre-determined battery of tests, utilizing a national norm for comparisons with a focus on within child characteristics (Reschly & Ysseldyke, 2002). Often times these batteries of tests were administered in the absence of any focused assessment question(s) which investigated the

alterable variables of the defined problem. Although the tests were technically sound for the purpose of identifying and categorizing students to pre-set disability criteria, data derived from their administration failed to consistently provide practitioners with information about how to design and evaluate educational treatments that resulted in consistent increases in student achievement (Cronbach, 1975).

The rationale for the use of broad tests of achievement and aptitudes was eloquently presented by Cronbach (1957) where he emphasized the use of both correlational and experimental approaches of psychology to identify aptitude by treatment interactions (ATIs). In the years to follow, researchers attempted to merge the two approaches to identify ATIs but failed to consistently demonstrate the link to interventions. In response, Cronbach (1975) revisited his position and advised that practitioners utilize short run empiricism to identify effective treatments instead of using aptitude and achievement patterns to deduce proper treatments. The necessity of using inductive hypothesis testing to demonstrate what worked for individual students was apparent in light of the ineffectiveness of using correlational research to consistently match treatments with identified aptitude/achievement characteristics to increase student outcomes (Cronbach, 1975; Kavale & Forness, 1999). Despite these findings, nationally norm referenced tests assessing aptitudes and achievement have continued to be the primary source of data used to determine special education eligibility in spite of a host of criticisms and a lack of treatment validity (Gresham & Witt, 1997). A primary reason for the continued reliance on traditional tests is connected to the administrative requirements of a categorical approach that necessitate these data (Reschly, Tilly, & Grimes, 1998).

Instead of focusing assessment on within-child constructs, such as processing, and the discrepancy of intelligence and achievement to determine eligibility, some have suggested that assessment investigates alterable variables, such as instruction, curriculum, observable student skills, and student response to intervention to make entitlement decisions (Deno, 1989; Heartland Area Education Agency, 2002; Howell, Fox, & Morehead, 1993). Research and discourse has increasingly focused on functional academic assessment tools and the potential advantages of utilizing a problem solving, interventions-based model. Practitioners using functional assessments can key in on student needs and response to intervention instead of a refer-test-place model emphasizing disability labels, ATIs, and placement in special education classrooms (Reschly, Tilly, & Grimes, 1998). A reliable and valid assessment system that has been demonstrated to provide teams with data to answer questions about a student's level of achievement and how a student responds to academic interventions is CBM (Shinn, 1989; Shinn, 1998).

Curriculum-Based Measurement

CBM is a set of standardized procedures used to present practitioners with a data base to assist when making a variety of educational decisions about basic skills in reading, math, written expression, and spelling (Deno, & Mirkin, 1977; Deno, 1989; Shinn & Bamonto, 1998). The assessment stimuli used with CBM procedures (i.e., administration rules and scoring directions) consists of short assessments, usually 1-3 minutes in length and are referred to as probes. The probes used with the procedures are constructed from the skill being assessed. The probes are sometimes sampled from the local curriculum but are increasingly being manufactured in pre made generic probe sets

that are constructed to lessen the variation in difficulty level (Hintze & Christ, 2004).

When using CBM procedures in reading, students are presented with a grade level passage that is approximately 250 words in length and are read a standardized set of directions instructing them to read for 1 minute. At the end of 1 minute, the student is asked to stop reading. This assessment results in an outcome of words correct per minute (wcpm) which is arrived at by summing the number of words read and subtracting the number of errors. Errors are documented when the student fails to pronounce the word after three seconds, substitutes a word, mispronounces a word, reverses a word, or omits a word (Shinn, 1989). Self corrections completed within three seconds are counted as correctly read words. The final score for each probe contains the number of wcpm and the number of errors committed.

CBM data are primarily used to provide information about the screening of students in need of further assessment, to define problems in the local educational context using specific target behaviors, and to evaluate the effectiveness of academic interventions (Deno, 1989). CBM allows educators to accomplish these goals by providing both summative and formative data that can be used to answer specific questions concerning the target behavior within the structure of the problem solving model (Deno, 1989; Heartland Area Education Agency, 2002). Data are described as summative when a score is representing a student's level of the measured construct or behavior at a specific point in time, as is seen in pre- and post-tests. Formative assessment consists of repeated measurements over time and when analyzed as a group, presents data about student growth or learning rate (Tawney & Gast, 1984). Curriculum-based measurement can provide data, in the form of wcpm that can be used in both a

summative and formative manner, to allow practitioners to make educational decisions about a variety of assessment questions. This contrasts most traditional published norm referenced tests that only provide data that are summative (Shinn & Bamonto, 1998).

Data collected using CBM can address the two primary questions of the NASP (2003) entitlement recommendations about student level of achievement and student response to intervention. A qualification of tests being used to present teams with information to make entitlement decisions is that they are reliable and valid for the purposes in which they are used. Expert judgments have identified several criteria for the reliability of tests for certain decisions. A common estimate for a sufficient reliability coefficient for individual assessment is a minimum of .80 (Sattler, 2001). However, general criteria may not be accurate given the context and data characteristics derived from the assessment. Adequate reliability coefficients for CBM are widely documented when making decisions about the relative standing of individuals and ranged from .89-.97 (Marston, 1989; Tindal, Germann, & Deno, 1983; Tindal, Marston, & Deno, 1983). However, disagreement exists about the number of data points needed to estimate a student's rate of learning (i.e., slope) with recommendations ranging from 8-16 data points (Hintze, Owen, Shapiro, & Daly, 2000).

Since CBM is a set of procedures, the assessment stimuli (i.e., probes) are not standardized and there exists a variety of sources to construct probes ranging from generic probe sets to using Shinn's (1989) recommendations for creating probe sets from the local curriculum. CBM, as implied in its name, was developed with the intention of sampling from the local curriculum to generate measurement probes to ensure a high level of content overlap (Shinn, 1989). However, various studies have demonstrated that

similar decisions were arrived at when using different probe sets implying that probes constructed from the local curriculum are not necessary to answer questions concerning student progress (Fuchs & Deno, 1994; Powell-Smith & Klug, 2001).

Hintze and Christ (2004) investigated the effect of probes controlled for difficulty on the standard error of the slope and determined that controlled probes reduced error in the slopes used to evaluate progress. Results showed that passage variability undermined the measurement capabilities of CBM and that probe difficulty was a significant source of error in obtaining an accurate estimation of a student's learning rate or slope. The results of these studies support the use of generic reading probe sets that are constructed to control for probe difficulty by limiting the variability of scores due to passage difficulty, as opposed to using probe sets sampled from basal texts (i.e., the local curriculum).

Decisions Made in a Problem Solving Model using CBM

Deno (1989) emphasized the importance of the marriage of assessment and evaluation and how this can help practitioners answer questions focused on alterable variables in the interaction between the educational environment and the student. The development of CBM was instrumental in providing standardized assessment procedures that presented data to practitioners to make educational decisions about a variety of questions rooted in the problem solving model (Tilly & Grimes, 1998). The problem solving model consists of four parts: 1) problem definition; 2) plan development; 3) plan implementation; and 4) plan evaluation (Bergan, 1977; Heartland Area Education Agency, 2002). Each of these steps contextualizes a series of questions to be answered by

a variety of assessment methods depending on the severity of the problem and the importance of the decision being made.

Problem Identification, Problem Validation, and Problem Definition

Steps that accompany the problem definition stage include problem identification, problem definition, and problem validation (Deno, 1989; Heartland Area Education Agency, 2002). A majority of the data used to make these decisions come from local normative data using CBM. Traditionally, these data are collected three times over the course of a school year in the fall, winter and spring. Each student receives a score calculated by taking the median wcpm score of three probes during each testing. These data present school personnel with a list of wcpm scores and corresponding percentile rank from the lowest to the highest performing student in the building.

The first decision that is made is to determine if a problem exists. The data used to make this decision comes from grade-level norms and allows for a relative comparison of individuals in the local context. This analysis can be used to identify the low performing students that may need additional instruction. This same data is also used to validate the nature of the problem. This is accomplished by looking at patterns in the grade-level data to rule out classroom or curriculum related problems. For example, if a school had four classrooms and one classroom accounted for a majority of the low performing students at the middle or end of the year, individually referring students would not solve the problem. If all four classrooms were low then the appropriateness of the curriculum would need to be investigated. When these explanations for low student performance are ruled out, screening data (i.e., student and local norm CBM scores) are used to define the problem. To define the problem, the student's score is compared to

peer scores and/or selected criterion to arrive at a discrepancy. The size of the discrepancy is used to gauge the intensity of the problem and to influence the amount of resources allocated to intervention.

CBM has been validated to answer questions concerning the relative standing of individuals by comparing student wcpm scores (Marston, 1989; Tindal, Germann, & Deno, 1983; Tindal, Marston, & Deno, 1983). However, no research could be located about the reliability of the baseline screening data used to represent a student's level of oral reading fluency (ORF). Confusion concerning how many probes to use to define an individual's level of ORF can be observed by the varying approaches that have been recommended to collect data representing a student's level of reading fluency (i.e., baseline point). Currently, there is no consensus on the number of probes needed to establish a student's baseline point with some recommending the use of the median score from three probes administered over one day and others saying to use the median score of nine probes over three days (Fuchs, 1989; Marston, 1989; Shinn, 1989).

Problem Analysis and Intervention Development and Planning

Questions asked in the problem analysis stage of problem solving revolve around why the problem is occurring and what instructional, curricular, and environmental alterations are most likely to enhance student learning. These decisions are often made by multidisciplinary teams that are presented with assessment data from a variety of sources, of which CBM could be one. Teams also investigate the setting, participants, and the amount of resources that can be used in carrying out the proposed intervention.

The assessment data collected during the problem analysis stage provides data to support inferences about why the identified problem is happening and what needs to be

done to remedy it (Bergan, 1977). The team uses these data to construct a plan. Instrumental activities that are completed in this stage of problem solving include answering assessment questions, setting goals, specifying a time for the team to meet to review progress, and determining decision rules for altering the intervention (Deno, Fuchs, Marston, & Shin, 2001; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Heartland Area Education Agency, 2002). A meta-analysis by Fuchs & Fuchs (1986) showed that the systematic implementation of a progress monitoring system significantly increased student progress .5 standard deviations over progress monitoring without it. The systematic components included the use of graphs, the establishment of goals and the definition of a goal line, and the use of either a goal-oriented or a treatment-oriented approach to evaluating intervention success (Fuchs, 1989).

An important activity during the intervention planning aspect of this stage is the development of a goal line and the corresponding growth rate to be documented on equal interval graphing paper. There are several strategies used to set the slope of goal lines (Fuchs, 2002; Heartland Area Education Agency, 2002). One strategy is to use school norming data and consider the slope of students at particular percentile ranks who perform at specified levels, (e.g., approximate percentile rank of low achieving, average, and above average students). Another is to use the predetermined growth rates published in the literature (Deno et al., 2001; Fuchs et al., 1993). Also, in using norming data, practitioners can identify the target student's discrepancy from peers and set the goal to be equal to where an average peer would be expected to be in a set period of time.

There are two procedures that are often used in the evaluation of the effectiveness of interventions, the goal-oriented approach and the program-oriented approach (Fuchs,

1989). With a goal-oriented approach to decision making, the practitioner compares the slope of the intervention with the slope of the goal line. If the slope of the intervention is not going to meet the goal at the end of the intervention period, alterations may be made to the intervention. When the slope resulting from the intervention is on course to meet or exceed the goal, the intervention is continued. In a treatment-oriented approach, there is no goal line and the practitioner decides on whether to continue with an intervention by comparing the obtained slope from different interventions. Fuchs (1989) recommended the use of a goal-oriented approach to decision making as it was easier to train teachers to conduct and resulted in more reliable decisions.

Another aspect that affects how decisions are made concerning the slope of student achievement is how the slope is calculated. There are three methods commonly used: an ordinary least-squares regression method, the split-middle technique, and visual analysis (Shinn, Good, & Stein, 1989). Investigations of these methods have demonstrated that using ordinary least-squares regression to calculate the slope is more accurate than the use of the split middle technique or a visual analysis (Shinn, Good, & Stein, 1989). The visual analysis of data points to draw a line of best fit is the least time consuming approach but often leads to unreliable decisions. The split-middle technique involves splitting the collected data in two equal halves, identifying the median of each half, and drawing a line through the two points to determine the slope (Stage, 2001). Although this is more systematic than a visual analysis, it does not as accurately depict student growth relative to the line of best fit calculated by computing the slope using a multiple regression equation. The use of the ordinary least-squares method, although the

most reliable, becomes problematic when data is not entered into a computer for analysis and when outliers exist in a limited data set.

Before an intervention is implemented, a variety of progress monitoring procedures should be defined that have been demonstrated to increase student response to intervention (Fuchs, 1989). These include the measurement strategy (who does it, what will be used, measurement conditions) and a decision making plan (frequency of data collection, how data will be summarized, number of data points before the analysis, and decision rules) (Heartland Area Education Agency, 2002). These practices set the foundation for the data-based evaluation of the intervention. One recommended procedure to evaluate interventions is to use a goal-oriented approach comparing a student's slope to the slope of the goal line (Fuchs, 1989).

Program Evaluation

The use of CBM with repeated measures provides a data set to educators with formative data to identify student learning rates over short periods of time (Deno & Mirkin, 1977; Fuchs, 1989; Fuchs & Fuchs, 2002; Hintze et al., 2000). A major distinction between traditional norm referenced assessments and CBM is the capability for practitioners to use CBM data in an inductive framework over the course of a short period of time, (i.e., 4 - 8 weeks). Traditional methods of program evaluation are largely summative in nature and consist of a pre- and post-test done on an annual basis. This does not allow teams to determine if instructional programming is effectively increasing a student's rate of learning over short periods of time. CBM data can be used by practitioners to measure how a student responds to academic interventions.

The formative assessment of interventions is a crucial step in problem solving and practitioners need to be aware of when decisions about slope, or learning rate, can be confidently made. Predictions for how many data points are needed to make a reliable and valid decision concerning the stability of the slope are mixed with estimates from 8-10 data points over 4 - 5 weeks (Hintze et al., 2000; Shinn, Good, & Stein, 1989) and 16 data points over 8 weeks (Hintze et al., 2000). Hard and fast rules are difficult to set as the confidence needed in the decision will be dependent on, and increase with, the importance of the decision being made (Heartland Area Education Agency, 2002). Other research-based progress monitoring practices include the use of probe sets controlled for variability in passage difficulty and the use of a regression line to calculate a student's rate of learning (Fuchs, 1998; Stage, 2001).

The number of data points needed before the slope can be confidently and accurately reported varies from study to study, and it likely depends on the characteristics of the probe set, age of the child, and method used to calculate the slope, but the research converges on a minimum of eight data points (Fuchs, 1989; Hintze et al., 2000). If CBM data were collected twice weekly, an intervention would need to be implemented for a minimum of one month to confidently make a decision about whether to change, modify, or continue the intervention. Unfortunately, if the intervention needs to be altered, the use of CBM data does not inform the team what to change. To gain this information, the team will have to revisit the problem analysis stage and discuss why the intervention was unsuccessful, construct a new plan, and modify or change the intervention (Bergan, 1977). Another decision at this point could be to discontinue, or fade, the intervention because the student met their goal (Bergan, 1977).

Reintegration

The goal of special education is to remediate educationally relevant skill deficiencies so that students can benefit from instruction in the general education setting (Powell-Smith & Ball, 2002). It has been estimated that approximately 36-40% of special education students are performing at a level commensurate to peers served without specialized instruction and would be good candidates for reintegration (Shinn, Habedank, Rodden-Nord, & Knutson, 1993). Once a student receiving special education services has evidenced the needed skills to meet the requirements of task demands in the classroom, then reintegration should be considered (Allen, 1989; Shinn et al, 1993).

The process of reintegration follows the same process of problem solving as previously discussed. The student is identified as having skills not significantly different from peers and the team investigates if acceptable performance and skill growth can be maintained in the general education setting without specialized instruction (Powell-Smith & Ball, 2002). A plan for reintegration is decided upon where supports are faded while progress is continually monitored to demonstrate that the student is benefiting from instruction in the general education setting.

After the reintegration plan is implemented (after a brief amount of time receiving instruction exclusively in the general education setting), the student's progress of skill development and performance on classroom objectives is investigated. The team uses data to decide if the student can receive a free and appropriate education without specialized instruction. Data used for making these decisions include the student's relative comparison to peers, a comparison of the student's slope or skill progress to

peers, and the student's performance on the task demands of the general education classroom environment (Powell-Smith & Ball, 2002).

Summary of CBM Applications to Problem Solving Model Decisions

CBM data are used to make a variety of important educational decisions within and across the problem solving model including a) problem identification, validation, and definition, b) problem analysis and the development of interventions, c) program evaluation, and d) reintegration. Given the importance of these data in influencing these decisions, it is critical that CBM procedures yield data that are reliable, valid, sensitive, and dependable so that appropriate decisions are made. While researchers have investigated the psychometric properties of CBM, there are limitations with this research-base, especially when inferring that reliability estimates using classical test theory generalize to within-student decisions.

Various Approaches to Estimating the Reliability of CBM

Reliability is a term that is frequently used in psychology but one that differs slightly depending on the definition. Two reliability models include the true score model of classical test theory developed by Spearman in the early 1900s and generalizability (G) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Both emphasize stability, but where classical test theory stresses the repeatability and consistency of measures, G theory focuses on the dependability or accuracy of the generalization of the test score based on the purpose and components of the testing situation.

Classical Test Theory and Reliability

Classical test theory, also known as true score theory, is the foundation of reliability theory and postulates that a person's observed score equals their true score plus

random or unsystematic error (Sattler, 2001). There is no mention of systematic error in true score theory. Furthermore, the appearance of systematic error would be indistinguishable from the random error estimate. Classical test theory is generally concerned with the relative standing of individuals, assumes that a hypothetical true score exists, and posits that forms of an assessment are parallel (Shavelson & Webb, 1991). In general, the reliability of an assessment is determined by dividing the shared or co-variance of test administration one and two by the product of multiplying the standard deviation of the test scores of test one and test two. The reliability coefficient represents the ratio of the variance of the true score to the variance of the observed score. Since the true score can not be known, reliability coefficients are estimates (Sattler, 2001).

There are four main types of reliability: test-retest reliability, alternate form reliability, internal consistency reliability, and interrater reliability. Test-retest reliability is represented by the coefficient of stability and is arrived at by administering the same test to a group of people on two separate occasions, usually close in time. Test-retest reliability coefficients are calculated by comparing the relative standing of individuals across the same, or parallel, forms. Alternate form reliability is a measure of the similarity of two forms of a test. For forms to be considered parallel, they must have the exact same difficulty level, a feat nearly impossible to achieve. Internal consistently reliability refers to the consistency of the instrument within itself, while interrater reliability is concerned with the consistency across different raters when assessing a behavior, trait, or construct (Sattler, 2001).

Criticisms and Limitations of Classic Test Theory

Traditional views of reliability have long been criticized by behaviorists with specific concerns about judging the quality of an assessment instrument through the consistency or stability of its scores and the supposed reality of a true score (Hartmann, Roper, & Bradford, 1979). Silva (1993) explains,

> "Behavioral assessors reject the interpretation of the observed score as consisting of true value plus measurement error, or, more concretely, they reject the interpretation of that "true" value as something consistent and stable. Taking consistency and stability as criteria of score quality implies an assumption that behavior is consistent and stable—an assumption made in trait theory" (p. 52).

Although this seems to strongly condemn the premise of reliability, behaviorists are more critical of the traditional view that error is random as they would see the stability, or instability, of a test as an interaction between the subject under investigation and the environment that would require explanation (Hartmann, Roper, & Bradford, 1979).

Instead of attributing variation in scores to random sources of error, behaviorists would postulate that scores vary across administrations for some verifiable reason, such as the conditions of the assessment setting, level of the measured attribute or skill, and/or the difficulty of the item or item set. A major drawback of classic test theory is its inefficiency to compartmentalize what was reported as random error. This does not allow for the identification and investigation of more than one source of error at a time.

Other criticisms and limitations of true score theory revolve around its rigidity for allowing only relative comparisons and the assumption that measurements are parallel (Shavelson & Webb, 1991). Although some assessment instruments are meant solely for

comparing an individual to a group, other assessments seek to investigate an individual's performance against himself or herself, without being compared to a group. For these types of assessments, classical test theory and traditional reliability coefficients are not appropriate because of their reliance on rank ordering and the assumption that test forms are parallel. For example, if two forms are given, one easy and one hard, and they rank order the sample similarly, even though the scores of the individual test takers were significantly different, a high alternate form coefficient would be obtained.

Generalizability Theory

In response to these criticisms, Cronbach et al. (1972) developed an alternative method to estimate reliability through the use of generalizability (G) theory. G theory addresses the aforementioned limitations of classical test theory. It also provides a more focused way of investigating the dependability of a measure and how inferences about a behavior or construct generalize to the defined universe constructed by the assessment conditions. In essence, G theory is not based on the traditional assumption that reliability and validity are separate, but assumes reliability and validity both fall on the same continuum of dependability (Silva, 1993). When students are administered a test, the examiner is not necessarily interested in that particular score unless that score is dependable. Shavelson and Webb (1991) emphasize that the score an examiner is after is one that would be observed from the student under multiple testing sessions under a similar testing environment. Inherent in this view is that scores will differ from administration to administration due to a variety of reasons, some of which could include the administrator, occasion, setting, and test forms. One application of G theory is to

identify and estimate the various sources of error that cause inconsistencies in the generalization of test scores.

Coefficients using G theory are obtained through the investigation and estimation of the variation in scores including the universe score or person facet and the other identified facets identified by the researcher. Shavelson and Webb (1991) explain:

> "…just as ANOVA partitions an individual's score into the effects for the independent variables, their combinations (interactions), and error, G theory uses the factorial ANOVA to partition an individual's score into an effect for the universe-score (for the object of measurement), an effect for each facet or source of error, and an effect for each of their combinations" (p. 16).

This is accomplished through the use of a repeated measures ANOVA.

G theory can provide data that separately estimates multiple sources of error simultaneously. These data inform practitioners about sources of error that may need to be reduced to obtain a dependable score and provide information about the generalizability of scores for both relative and absolute decisions (Shavelson & Webb, 1991). Two types of studies can be used to accomplish these goals, generalizability (G) studies and decisions (D) studies. Shavelson and Webb (1991) elaborate:

> "The purpose of a G study is to anticipate the multiple uses of a measurement and to provide as much information as possible about the sources of variation in the measurement. A D study makes use of the information provided by the G study to design the best possible application of the social science measurement for a particular purpose" (p. 12).

To conduct a D study, the researcher must decide on the facets that will be used to define the universe. Also the purpose of the assessment and whether data will be used for relative or absolute decisions need to be identified. The information obtained concerning the error components of the G study can be used to investigate the combinations of assessment designs to maximize the dependability of the generalization. For example, if the person facet accounted for 60% of the variance in scores, more samples of the behavior would need to be taken to obtain a reliable score than if the person facet accounted for 80% of the variance. This allows for the construction of assessments and the design of administration procedures that can increase the efficiency in which practitioners arrive at a dependable score for the type of decision they are trying to make. This is similar to how the Spearman-Brown prophecy formula informs about the length of a test to make relative decisions (Shavelson & Webb, 1991).

G theory calculations can be used to inform practitioners about both relative and absolute decisions. This is especially relevant to CBM data as they are used in a problem solving model to address both of these types of decisions. Relative decisions investigate how a student compares to other students in the local educational environment. Absolute decisions compare how the student's score on one form or occasion compares to his or her score on another form or occasion, an intraindividual comparison. Although the application of classical test theory can offer insight into the reliability of CBM data for making relative decisions, until recently researchers failed to investigate the reliability of CBM for making absolute decisions (Hintze et al., 2000; Hintze, Christ, & Keller, 2002). The investigation of this type of decision is critical if CBM is going to be used to make absolute decisions about individual performance over short periods of time.

Relative and absolute reliabilities have the potential to vary little or dramatically depending on whether the variation in scores is solely due to facets that interact with the object of measurement or if facets exist independent of the object of measurement. This becomes more of a concern with measures that are used repeatedly whether it be observations or alternate test forms (Hintze et al., 2000). An example will be used to illustrate this point.

In a hypothetical example, one group of students are given three reading probes of various difficulty levels, one is highly difficult, one is at a medium difficulty level, and one is easy. Another group is given three probes that are all closely matched and consequently have a similar level of difficulty. Assuming that each probe similarly rank ordered student oral reading fluency, the reliability coefficient as reported in classic test theory would provide similar estimates for both sets of probes. G theory's relative coefficient, the generalizability (G) coefficient, would give similar estimates concerning relative decisions for both groups. This is because the rank ordering of scores negates the variation in student scores caused by different levels of probes. However, for absolute decisions, how other students performed is irrelevant. The absolute reliability estimate reflects the variation in scores caused by the lack of homogeneity across probes, in whereas relative reliability would not. The group that received the easy, medium, and hard probe would exhibit a high level of relative reliability and lower level of absolute reliability. The group receiving probes that were all of medium difficulty would exhibit a high level of relative reliability and a similarly high level of absolute reliability. Provided the forms were parallel, as assumed in classical test theory, the reliabilities would be virtually equal.

G theory calculations can be used to investigate the dependability of a measure for both relative and absolute decisions, to partition out multiple sources of error, and to provide information on the assessment characteristics needed to arrive at a dependable score. Thus G theory studies are a logical choice for determining the reliability of CBM. Unfortunately, researchers investigating the psychometric characteristics of CBM have largely ignored this methodology until two recent studies, one involving reading (Hintze et al., 2000) and the other math (Hintze, Christ, & Keller, 2002).

A Review of the Studies Investigating the Reliability of CBM

Two studies (Tindal, Germann, & Deno, 1983; Tindal, Marston, & Deno, 1983) and/or one book chapter (Marston, 1989) are often referenced when reporting on the reliability of CBM reading passages. For the purposes of this paper, only the methodology and corresponding coefficients for reading from these studies and book chapter will be discussed. Neither of the studies used G theory. Both used true score theory to investigate test-retest, alternate form, and interrater reliability (Tindal, Germann, & Deno, 1983; Tindal, Marston, & Deno, 1983).

The use of classical test theory was appropriate for the stated purpose of the Tindal, Germann, Deno (1983) study which was done to investigate the reliability of CBM for norming. This study reported on test-retest and alternate form reliability. The obtained test-retest reliability estimate of .97 was based on 30 fifth-grade students with a 2-week interval. This validated that the same CBM passage could reliably rank order students. The alternate form reliability estimate of .94 was based on 110 fourth-grade students who were given two forms during one administration session. This information,

when interpreted for relative decisions, validated that when given two alternate probes, each probe similarly rank ordered students.

The study completed by Tindal, Marston, & Deno (1983) defined the purpose of their study to examine the reliability of using CBM for repeated measurement. Three third-grade level probes were administered to 566 students from grades one through six. The estimated reliability coefficient for interrater reliability was .99. The test-retest coefficient estimated at .92 was obtained after a 10-week delay and the alternate form coefficient estimated at .89 was obtained after a 1 week delay as reported by Marston (1989). These reliability estimates are consistent with recent evaluations investigating the reliability of CBM for relative decisions (Hintze et al., 2000).

Implications for CBM

The data from the Tindal, Marston, and Deno (1983) study reported that the use of repeated measurements were reliable. What was not reported was for what purposes CBM data was and was not reliable for. Given their use of classical test theory methods and procedures, Tindal et al. (1983) showed that CBM was reliable when rank ordering students. Their data did not provide information about the use of CBM for absolute or intraindividual purposes such as progress monitoring. In essence, there was an absence of reliability data about using CBM to make intraindividual or within-subject decisions until Hintze et al. (2000) approached the question using G theory. In the meantime CBM was widely reported as a reliable measure for formative assessment purposes for over 15 years.

Statement of the Problem

CBM was originated with the intention of presenting special education teachers with an assessment system that would present data that would be useful for a variety of purposes (Deno, 1989). A primary strength was the quick and efficient manner in which CBM procedures could be used to formatively assess student progress over short periods of time. By the mid 1980s, the uses of CBM data were expanded to include creating local norms and making screening and eligibility decisions based on the relative achievement of students. In 2004, CBM data continues to be used for the aforementioned purposes but are also being used to measure students' response to intervention. Student RTI data are now being used as a primary indicator when making special education eligibility decisions (Heartland Area Education Agency, 2002). Other uses of CBM now include validating the successful reintegration into general education, conducting error pattern analyses, instructional grouping, and its use in brief hypothesis testing (Howell, Morehead, & Fox, 1993; Shinn et al., 1993).

While the research literature abounds with references to the reliability and validity of CBM, a majority of the psychometric data for the reliability of CBM in reading was collected, analyzed, and reported during the early 1980's using classical test theory (Tindal, German, & Deno, 1983; Tindal, Marston, & Deno, 1983) and summarized in an often cited book chapter by Marston (1989). Although the uses and purposes of CBM have expanded, it is possible that practitioners still justify the use of CBM for high stakes decisions based on the reliability estimates of research reports published over 20 years ago using true score theory (Tindal, German, & Deno, 1983; Tindal, Marston, & Deno, 1983). The research community has yet to revisit the reliability of CBM even though its

usage has been expanded to make entitlement decisions using both relative and absolute (i.e., intraindividual) data.

The selection of a data point to define the problem can significantly affect decision making especially when systematic decision rules in progress monitoring are not used. Deno, Fuchs, Marston, & Shin (2001) defined "reasonable" growth rates for a third-grade student as an increase of one word correct per minute per week. It is important for there to be a limited amount of error in the baseline point if goal lines and decision making rules are implemented over short periods of time. Depending on the decision rules used, the amount of data points used, and the way the data points are analyzed, a baseline over-estimation of six wcpm could lead to the assumption that an intervention was unsuccessful because the student did not achieve the goal. This would result in a decision to change instruction as the goal trajectory would need to be increased to meet the goal. If the baseline was appropriately set, the student would have made the goal and demonstrated that the instructional practices implemented were appropriate.

Hintze et al., (2000) demonstrated through two studies utilizing G theory that they could make a highly reliable (.90 and .82 g-coefficients respectively) decision using 16 data points collected over 8 weeks to measure student improvement (i.e., slope). This information would suggest that teams could confidently assess progress and make reliable decisions about the trend of student learning, lessening the importance of using a goal line for comparison. However, this would also assume that two data points were collected per week to make the decision at eight weeks, the slope of peers in the local educational context was known, and that the teachers were using regression lines and not a split middle or visual analysis technique to determine the slope. In addition, Fuchs

(1989) has stressed the importance of goal lines and decision making rules to increase the accuracy of decisions concerning student response to intervention. Adhering to recommendations for progress monitoring as delineated by Fuchs (1989) and field-based problem solving models, e.g., Heartland Area Education Agency (2002), a dependable baseline and definition of the problem is critical to the evaluation of future progress.

Purpose of the Study

Educators and researchers have focused on using CBM data to answer two types of questions. One type of question is concerned with a student's discrepancy in the level of academic achievement as compared to peers. The second type of question investigates a student's response to intervention (RTI) when presented with an intervention matched to his or her current instructional needs. These data are combined and used to make decisions regarding eligibility for special education.

Although the slope is the primary determinant in formative assessment, when making decisions over short periods of time (4 to 8 weeks), an accurate and stable summative score is paramount to the foundation of formative assessment. This estimated level of performance is vital to the problem definition stage of problem solving and impacts goal setting, the evaluation of student growth (i.e., goal lines), and the outcomes of arbitrary decision making rules (3 or 4 point under the goal line rule). At the present time, no peer reviewed research studies were located investigating the potential impact of error in the definition of a reading fluency problem and its potential effects on decision making. The professional literature has primarily focused on using CBM data to estimate the trend, or slope, of skill growth and to provide practitioners feedback about the success of implemented interventions. The absolute reliability has been investigated for making

decisions about slope (Hintze et al., 2000). However, data concerning the confidence of the defined level of performance and the corresponding standard error of measurement (SEM) of a student's observed score has yet to be investigated.

The purpose of the current study is to use G theory to conduct both a generalizability (G) study and a decision (D) study. The goal of the G study will be to provide scores concerning the estimated variance components of each facet included in the investigation along with the percentage of total variance for each. Facets included are student skill level, probes, and sequence. The purpose of the D study will be to determine the optimal number of CBM probes needed to confidently make a variety of educational decisions. The decisions investigated will include the number of probes to make relative decisions and the number of probes needed to make the absolute decision concerning the identification of a stable baseline point. In addition, an index of dependability will be obtained for a variety of probe combinations that could be used to define the baseline point, (i.e., 3 probes, 5 probes, 7 probes, and 9 probes). To place the generalizability coefficients in a proper context, SEMs will be reported.

It is hypothesized that probe difficulty will represent a significant percentage of the error variance. Subsequent D studies will be conducted by removing probes that vary from the mean of the probe set to attempt to reduce the percentage of error variance due to probe difficulty, to reduce the SEM that accompanies the summative data point, and to increase the dependability estimates.

Research Questions

1. What percentage of the variance in scores is due to the following facets: person (p), items (i), sequence (s), and the residual p x i,e?

2. What is the generalizability coefficient and SEM for relative decisions about rank ordering students using one probe and the average of three probes?
3. What is the index of dependability and SEM for the absolute decision about a student's baseline score using the average of three, five, seven, and nine probes?
4. How are the aforementioned questions affected when items that deviate from the average score of the probe set by +/-15 wcpm, +/-10 wcpm, and +/-5 wcpm are removed?

CHAPTER II

METHOD

Participants and Setting

Study 1:

After obtaining permission to conduct the study, all students from two third-grade classrooms in a rural Iowa school district were given an informed consent form to take home to their parents. Only students whose parents signed and returned the forms were included in the study. Out of 27 possible subjects, the sample included 14 students. The sample included 3 (21%) males and 11 (79%) females, ages ranging from 8-10 years old. One hundred percent of the sample was Caucasian, one student (7%) received free and reduced lunch, and none of the students received special education services in the area of reading. Permission to conduct the study was obtained from the district superintendent, building principal, and the University of Tennessee Institutional Review Board. In addition, parental permission and student assent was required for all participants (See Appendix A for parental consent, student assent, and school permission forms).

Study 2:

After obtaining permission to conduct the study, all students from four third-grade classrooms in two rural Iowa school districts were given an informed consent form to take home to their parents. Only students whose parents signed and returned the forms were included in the study. Out of 57 possible participants, the final sample included 37 Caucasian students, 12 (32%) males and 25 (68%) females, ages 8-10 years old. Four participants (11%) received free or reduced lunch. Two (5%) students received special education services and both of their individual education plans (IEP) targeted reading

skill deficits. Permission to conduct the study was obtained from the district superintendent, building principal, and the University of Tennessee Institutional Review Board. In addition, parental permission and student assent was required for all participants.

Materials

Study 1 and study 2:

A probe set consisting of 20 third-grade level CBM oral reading fluency probes was used for the study and was obtained from the dibels.uoregon.edu website (Good, Kaminski, & Dill, 2002). The probe set used was not constructed from either of the schools' curricula and was "generic" with a focus on minimizing variation in probe difficulty with Spache readabilities ranging from 2.8-3.1 (Good & Kaminski, 2002). For both study 1 and study 2, the 20 probes were arranged in four packets of five probes for each day of data collection. In study 1, each student received all of the probes in the same order over the four day period. In study 2, the order of the 20 probes was randomized for each student using a random number table. Other materials used included a stopwatch, tape recorder, clipboard, pencil, and administration directions.

Procedures

Preliminary activities. Two weeks prior to collecting data, teachers provided each student with a consent form to be taken home. On the day before data collection was to begin, the students were given a brief explanation about the procedures and their assent was solicited. All students whose parents provided consent also provided assent. All participating students were to be tested each of 4 days. If a student missed a day, he/she was tested on the 5th day of the week. If a student missed two days, he/she was dismissed

from the study.

CBM reading administration. Testing was conducted in a small room that was generally used for counseling and testing purposes. Students were escorted from the classroom to the assessment area where they sat across from the examiner and were read the following directions, "When I say 'please begin' start reading aloud at the top of this page. Read across the page. Try each word. If you come to a word you do not know, I'll tell it to you. Be sure to do your best reading. Are there any questions? Please begin." The student began reading and errors were recorded. After 1 minute the student was asked to stop and to go to the next page. The directions were read prior to the first probe administered at the beginning of each day. To prompt the student to begin on the remaining probes the experimenter said, "please begin." The stopwatch and scoring probe were both out of the sight of the student. When the student was finished reading the probes he/she was escorted back to the classroom and the next student was selected.

Scoring directions. Errors were documented when the student failed to pronounce the word after 3 seconds, substituted a word, mispronounced a word, reversed two words, and omitted a word. If a student self-corrected an error within 3 seconds, the self-correction was counted as a correctly read word (Shinn, 1989). The final score for each probe contained the number of wcpm and the number of errors committed. Only wcpm were analyzed in this study.

Training procedures/interscorer agreement. All data were collected by the principal researcher who, prior to the study, was trained to administer and score CBM. Approximately 15% of the CBM administrations were taped recorded and checked for interscorer agreement by an independent scorer. To calculate the percentage of

agreement, the number of agreements was divided by the number of agreements plus disagreements and multiplied by 100. Interscorer agreement data was calculated from 151 of the 1020 (15%) probes administered. The mean number of agreements was 98% with a range from 90%-100%.

Data Analysis

The data were analyzed using Generalizability (G) theory to conduct both a Generalizabiltiy (G) study and a Decision (D) study. Facets under investigation in studies 1 and 2 included the object of measurement, or persons (p), probes/items (i), and residual error (pi,e). In addition, study 2 also investigated the variance caused by testing effects. Both examinations used a single facet, crossed design to arrive at the generalizability coefficients and corresponding SEMs for relative and absolute decisions using summative wcpm data. All of the calculations using G theory were based on formulas described by Shavelson and Webb (1991) and/or Brennan (2001).

The G Study. A G study was used to investigate the variation of wcpm scores due to student skill (persons), the difficulty of passages or probes (items), and unaccounted sources of error (residual). Statistical Package for the Social Sciences (SPSS) was used to run a repeated measures ANOVA to obtain the sums of squares (SS) for each facet. The mean squares (MS) for each facet were obtained by dividing the SS by the degrees of freedom (df) for each facet. The resulting MS of the residual ($MS_{pi,e}$), which is equal to the estimated variance component of the residual ($\sigma^2_{pi,e}$), was used in Equations 1, 2, and 3 to arrive at the remaining estimated variance components (σ^2_i and σ^2_p).

$$\sigma^2_{pi,e} = MS_{pi,e} \quad (1) \qquad \sigma^2_i = (MS_i - \sigma^2_{pi,e}) \div n_p \quad (2) \qquad \sigma^2_p = (MS_p - \sigma^2_{pi,e}) \div n_i \quad (3)$$

The estimated variance components were summed and each individual estimated variance

component was divided by the sum of estimated variance components to arrive at the percentage of total variance for each of the three facets. This sequence of calculations was used in each of the G studies.

The D Study. Data about the estimated variance components for the person, item, and residual facets from the G study were used to conduct the D study. The purpose of the D study was to determine the optimal number of CBM probes needed to confidently make relative and/or absolute decisions. The outcomes of the D study resulted in a reliability-like coefficient and a SEM for both relative and absolute decisions depending on the number of probes placed in the analysis. The reliability-like estimate for relative decisions is referred to as the coefficient of generalizability (ρ^2) and the estimate for absolute decisions is called the index of dependability (Φ) (Shavelson & Webb, 1991).

One step to reaching the desired outcomes of the D study was to derive the estimated error variance for relative and absolute decisions. The *relative* error variance was obtained by dividing the estimated variance of the residual by the number of items proposed as seen in Equation 4. The *absolute* error variance was calculated by summing the products of the estimated variance component of items divided by the number of items and the estimated variance of the residual divided by the number of items proposed as seen in Equation 5.

$$(4)\quad \sigma^2_{Rel} = \sigma^2_{pie} \div \acute{n}_i \qquad\qquad (5)\quad \sigma^2_{abs} = \sigma^2_{i} \div \acute{n}_i + \sigma^2_{pi,e} \div \acute{n}_i$$

The generalizability coefficient and index of dependability were obtained using the relative and absolute error variances. The generalizability coefficient was derived by dividing the estimated variance component of persons by the sum of the estimated variance component of persons (Equation 6) and the relative error variance component

given the number of probes under investigation. The index of dependability was obtained by dividing the estimated variance component of persons by the sum of the estimated variance component of persons and the absolute error variance component (Equation 7).

(6) $\rho^2 = \sigma^2_p \div (\sigma^2_{p+} \sigma^2_{Rel})$ (7) $\Phi = \sigma^2_p \div (\sigma^2_{p+} \sigma^2_{abs})$

The last calculation in the D study was to arrive at the SEM for relative and absolute decisions given the number of items used. The SEM, when using a specified number of items, was calculated for relative decisions by taking the square root of the relative error variance component given the number of items used. The SEM for absolute decisions was achieved by taking the square root of the absolute error variance component given the number of items used.

CHAPTER III

RESULTS

Study One

The results summarized in Table 1 indicated the percentage of variance accounted for by person, item, and residual facets when using CBM procedures with the DIBELS third grade level probe set. Results showed that the largest amount of variation in scores, 69 percent, was attributable to the person facet (i.e., student skill). Item or probe difficulty accounted for 20 percent of the variance, and 11 percent of the variation was located in unaccounted sources of error. The estimated variance components for these facets were used for the D study investigating the coefficient of generalizability, the index of dependability, and their corresponding SEMs given the administration of 1, 3, 5, 7, or 9 probes. Results of the D study are reported in Table 2. The coefficient of generalizability ranged from .85 (SEM 11 wcpm) when one probe was administered to .98 (SEM 4 wcpm) when administering 9 probes. When using three probes, as is typical for creating local norms or screening students for further assessment, the coefficient of generalizability was .94 (SEM 6 wcpm). The index of dependability ranged from .68 (SEM 18 wcpm) with one probe to .95 (SEM 6 wcpm) when using nine probes to make a within-individual decision.

Study Two

Table 3 summarizes the percentage of variance accounted for by person (student skill), item (passage or probe difficulty), and residual (unaccounted error variance) facets. Results showed that the largest amount of variation in scores, 81%, was attributable to the person facet. Item, or probe, accounted for 10% of the variance, and 9% of the variation

TABLE 1: Estimates of Variance Components for Study 1

Sources of Variation	Sums of Squares	df	Mean Squares	Estimated Variance Components	Percentage of Total Variance
Persons (p)	179932.86	13	13840.99	686.31	69
Items (i)	54161.04	19	2850.58	195.49	20
Residual (pi,e)	28335.21	247	114.72	114.72	11
Total				996.52	100

TABLE 2: Decision Analysis for Study 1[1]

Sources of Variation	σ^2	$\acute{n}_i$ =	G Study 1	Alternative D Studies 3	5	7	9
Persons (p)	σ^2_p		686.31	686.31	686.31	686.31	686.31
Items (i)	σ^2_i		195.49	65.163	39.098	27.927	21.721
Residual (pi,e)	$\sigma^2_{pi,e}$		114.717	38.239	22.943	16.388	12.746
σ^2_{Rel}			114.717	38.239	22.943	16.388	12.746
σ^2_{Abs}			310.207	103.402	62.041	44.315	34.467
ρ^2			.85	.94	.97	.98	.98
Φ			.68	.87	.91	.93	.95
SEM ρ^2	+/- wpm		11	6	5	4	4
SEM Φ	+/-wpm		18	10	8	7	6

[1] Definitions of symbols in Table 2: $\acute{n}_i$ = number of items; ρ^2 = coefficient of generalizability; Φ = index of dependability; SEM = standard error of measurement; wpm = words per minute.

TABLE 3: Estimates of Variance Components for Study 2

Sources of Variation	Sums of Squares	df	Mean Squares	Estimated Variance Components	Percentage of Total Variance
Persons (p)	977471.60	36	27151.99	1350.08	81
Items (i)	125339.16	19	6596.80	174.23	10
Residual (pi,e)	102802.46	684	150.30	150.30	9
Total				1674.61	100

was located in unaccounted sources of error. Additionally, the amount of variance due to sequence (i.e., practice) effects was investigated. Results indicated that virtually none of the variability in wcpm scores was due to sequence effects.

The estimated variance components for these facets were used for the D study investigating the coefficient of generalizability and index of dependability and their corresponding SEMs given the administration of 1, 3, 5, 7, or 9 probes. Results of the D study are reported in Table 4. The coefficient of generalizability ranged from .90 (SEM 12 wcpm) when one probe was administered to .99 (SEM 4 wcpm) when nine probes were administered to make a relative decision. When using three probes, the coefficient of generalizability was .93 (SEM 7 wcpm). The index of dependability represents absolute decisions and ranged from .81 (SEM 18 wcpm) with one probe to .97 (SEM 6 wcpm) when using nine probes to make within-student decision (e.g., RTI).

Due to the random assignment of the administration of the 20 probes across students, practice and/or interaction effects were evenly distributed across probes allowing for the comparison of the average score for each probe. The average score for each individual probe was summed and an average score of 122.6 wcpm was calculated. This average was used to alter the probe set to include passages within a range of +/-15 wcpm, +/-10 wcpm, and +/-5 wcpm of the mean. G and D studies were conducted on each progressively homogenous set of probes to investigate the total percentage of variance due to person, items, and residual facets and provide reliability estimates and SEMs for both relative and absolute decisions.

TABLE 4: Decision Analysis for Study 2[2]

Sources of Variation	σ^2	$\acute{n}_i$ =	G Study 1	Alternative D Studies 3	5	7	9
Persons (p)	σ^2_p		1350.08	1350.08	1350.08	1350.08	1350.08
Items (i)	σ^2_i		174.23	58.08	34.85	24.89	19.36
Residual (pi,e)	$\sigma^2_{pi,e}$		150.30	50.10	30.06	21.47	16.70
σ^2_{Rel}			150.30	50.10	30.06	21.47	16.70
σ^2_{Abs}			324.53	108.18	64.91	46.36	36.06
ρ^2			.90	.96	.98	.98	.99
Φ			.81	.93	.95	.97	.97
SEM ρ^2	+/- wpm		12	7	5	5	4
SEM Φ	+/-wpm		18	10	8	7	6

[2] Definitions of symbols in Table 4: $\acute{n}_i$ = number of items; ρ^2 = coefficient of generalizability; Φ = index of dependability; SEM = standard error of measurement; wpm = words per minute.

Table 5 provides the percentage of total variance for the person, item, and residual facets when the probe set was altered to reduce passage variability. Results showed a reduction in the percentage of total variance due to items from 10% when using the original probe set to 1% when restricting the average score on any one probe to not exceed +/-5 wcpm from the average. Conversely, the amount of variance due to the person variance increased from 81% when using the original probe set to 89% when restricting the average score on any one probe to not exceed +/-5 wcpm from the average. The percentage of total variance accounted for by the residual factor was not significantly influenced by controlling for difficulty in the probes analyzed.

Table 6 presents the coefficient of generalizability, index of dependability, and their corresponding SEMs in response to altering the probe sets. Because altering probe difficulty would not influence the relative standing of individuals, the relative coefficients were not affected by decreasing the variability via controlling passage difficulty of the analyzed probe sets. The absolute coefficients increased in response to decreasing the variability in probe sets; however, this increase was limited when the coefficients for giving 5, 7, and 9 probes were calculated. The SEMs demonstrated a similar pattern with SEMs of 18, 10, 7, 6, and 5 wcpm when using 1, 3, 5, 7, and 9 probes, respectively, when using the intact DIBELS third-grade probe set. To investigate the affect of lessening the variability within the analyzed probe set, only probes within +/-5 wcpm of the probe set average were used. The SEMs of the altered probe set were 12, 7, 5, 5, and 4 wcpm when using 1, 3, 5, 7, and 9 probes, respectively. These results showed that controlling for probe difficulty would reduce the number of probes needed from 9 to 5 to achieve a commensurate level of error.

TABLE 5: Percentage of Total Variance with Altered Probe Sets

Sources of Variation	Percentage of Total Variance Probe Set	Percentage of Total Variance +/- 15 wcpm	Percentage of Total Variance +/- 10 wcpm	Percentage of Total Variance +/- 5 wcpm
Persons (p)	81	85.5	89	89
Items (i)	10	5.5	2	1
Residual (pi,e)	9	9	9	10
Total	100	100	100	100

TABLE 6: The Affect of Altered Probe Sets on ρ^2, Φ, and SEMs[3]

			Number of Items				
			1	**3**	**5**	**7**	**9**
Probe Set	ρ^2		.90	.96	.98	.98	.99
	Φ		.81	.93	.95	.97	.97
	SEM ρ^2	+/-wpm	12	7	5	5	4
	SEM Φ	+/-wpm	18	10	8	7	6
+/-15	ρ^2		.90	.97	.98	.98	.99
	Φ		.85	.95	.97	.98	.98
	SEM ρ^2	+/-wpm	12	7	5	5	4
	SEM Φ	+/-wpm	15	9	7	6	5
+/-10	ρ^2		.91	.97	.98	.99	.99
	Φ		.89	.96	.98	.98	.99
	SEM ρ^2	+/-wpm	12	7	5	4	4
	SEM Φ	+/-wpm	13	7	6	5	4
+/-5	ρ^2		.90	.96	.98	.98	.99
	Φ		.89	.96	.98	.98	.99
	SEM ρ^2	+/-wpm	12	7	5	4	4
	SEM Φ	+/-wpm	12	7	5	5	4

[3] Definitions of symbols in Table 6: ρ^2 = coefficient of generalizability; Φ = index of dependability; SEM = standard error of measurement; wpm = words per minute.

CHAPTER IV

DISCUSSION

There were three purposes of the current study. The first was to determine the percentage of variability in CBM scores due to student skill, probe difficulty, sequence/practice effects, and unaccounted sources of error. The second was to estimate reliability-like coefficients and a corresponding SEM for a summative data point given a specified number of probes to make relative and absolute (i.e., intraindividual) decisions. The third and final purpose was to investigate the magnitude to which altering the probe set's passage variability would reduce the SEM when defining a baseline point given a specific number of probes.

A goal of assessment is to be able to accurately estimate a student's true score on a relevant skill or construct to contribute to valid decision making. Results of study one showed that 69% of the variance in CBM scores was due to the person facet, 20% was due to probe variability, and the remaining 11% of the variance was due to unaccounted sources of error. Findings in study two showed that 81% of the variability was due to the person facet, 10% due to probes, and 9% due to the residual. Both studies strongly support the notion that CBM provides data that measures the oral reading fluency skills of the student as compared to some other variable(s) in the testing environment.

Consistent with reliability coefficients reported across the literature, the current studies reported that the coefficient of generalizability (used for relative decisions) ranged from .85-.99 and the index of dependability (used for absolute decisions) ranged from .68-.97 when making a decision about a summative point given a specified number of probes ranging from 1 to 9. Estimates of stability (e.g., coefficient of generalizabilty

and index of dependability) are generally helpful; however, they fail to translate the coefficient to the scores being used to drive decisions in the schools. To present practitioners with a numerical representation, in wcpm, of the amount of error associated with the administration of 1, 3, 5, 7, or 9 probes, the SEM for each was calculated. When using the standard probe set to estimate a student's baseline, an absolute decision, the SEMs ranged from 18 wcpm when using one probe to 6 wcpm when administering 9 probes in both studies. When making a decision about the relative standing of students, the SEMs ranged from 11 wcpm when using one probe to 6 wcpm when using 9 probes.

CBM was developed to present practitioners with a method to efficiently collect data to make educational decisions about basic skill development (Deno & Mirkin, 1977; Deno, 1989). In order to make decisions efficiently, practitioners must be presented with stable and accurate data (i.e., data with low levels of error). Hintze and Christ (2004) demonstrated that reducing passage variability within a probe set reduced the SE(b) and the SEE when using CBM procedures. Whereas the Hintze and Christ study limited the variability of the probe set through readability formulas, the current study reduced the variability of the probe set by including only those passages that had an average score within +/-15, +/-10, and +/-5 wcpm of the average probe set. When restricting the passages to +/-10 wcpm, the SEM was reduced to 13 wcpm when using one probe to 4 wcpm when using nine probes to address intraindividual questions about student level of performance. To address the call of Hintze and Christ to explore ways of determining the most efficient manner to collect CBM data, the current data suggests constructing probe sets with passages that deviate from the set average +/-5 wcpm and use five probes. This would put the SEM at 5 wcpm as compared to the current probe set that has a SEM of 8

wcpm when giving five probes and a SEM of 6 wcpm when administering 9 probes. In essence, using a probe set constructed using these guidelines would allow practitioners to more accurately and efficiently estimate student performance by decreasing the number of probes given from 9 to 5, while also reducing the error term.

Implications for Practice in a Problem Solving Model

The current studies investigated the psychometric characteristics of wcpm when using CBM procedures to estimate a summative point. A summative data point is used for several purposes in a problem solving model, such as problem identification, problem validation, problem definition, and defining a student's baseline level (Deno, 1989; Shinn & Bamonto, 1998). Depending on the characteristics of the testing environment, these decisions could be either relative or absolute in nature. Practitioner knowledge of whether or not data is being used in a relative or absolute capacity is needed to correctly identify the error term that should be used, with relative data having the smaller error term.

This reduction of error occurs because probe difficulty is nullified as a factor when all students receive the same probe(s). However, this decrease in error for making a "relative" decision requires that all students receive identical probes. For example, if yearly norms, (i.e., benchmarking) are used to make decisions regarding the relative standing of individuals and/or to define student discrepancy, then using identical probes across scores will have a lower SEM. However, if you are comparing a student's score to pre-existing normative scores, generic instructional standards, or cut off scores developed using *different probes*, the larger SEM associated with absolute decisions should be applied. The SEM associated with absolute decisions would apply when making within-student comparisons in order to assess RTI due to the administration of multiple alternate

forms. Furthermore, in each of these across probe comparisons (e.g., comparing current score to pre-existing norm, generic instructional standards, within-subject scores on pre- and post-tests), the error is applied to both scores. Thus, confidence in making accurate decisions with 3 probes across both comparison scores would require a difference of 20 wcpm (i.e., a SEM of 10 for absolute decisions).

A common decision made involving the rank ordering of students using CBM is the screening of students who may need extra instructional support. The data from the D study generally supports the Ardoin et al. (2004) recommendation to use a single probe for universal screening/problem identification to identify the possible pool of students in need of intervention. However, due to the large SEM of 12 wcpm around relative data points and the scarcity of school resources, further assessment using three (SEM of 7 wcpm) to five (SEM of 5 wcpm) probes would need to be administered to the group of students around the cut-off to be able to more accurately identify students who would receive more intensive support. This recommendation would hold true whenever using cut scores to make decisions about resource allocation, whether it be for an intervention program, determining instructional placement, or the estimation of a percentile rank to aid in eligibility decisions. Unfortunately, many of these decisions are not made when collecting grade level CBM data (i.e., the norm) when the reduced error term that accompanies relative decisions would be present. Therefore, the error associated with absolute decisions would need to be used when estimating a student's level of performance.

When validating and defining a problem, the practitioner defines the target student's current level of oral reading fluency and compares it to a standard, whether it is

based off of local norms or an expert judgment such as Shapiro's (1996) instructional recommendations or the Howell et al. (1993) criterion for acceptable performance (CAP). The discrepancy of wcpm between the student's performance and the standard defines the problem and validates the intensity of the recommended treatment. Depending on the size of the discrepancy, the SEM associated with the student score takes on more or less importance (e.g., if a student is 60 words discrepant, whether his or her true score is discrepant by 50 or 70 wcpm, the student needs intensive remediation). However, in today's climate of prevention and remediation, detecting small discrepancies is increasingly important which would necessitate decreased error terms around the data used to guide educational decision making.

Some of the previous decisions discussed could have been relative or absolute depending on the testing conditions. However, when CBM procedures are used to generate summative data to define a student's baseline point, as in progress monitoring, the error term associated with absolute decisions will always be used. A common example of this is when a summative score is used to anchor an aimline from which to evaluate a student's response to intervention. Recommendations for the number of data points needed to define a baseline point vary with some recommending using three data points and some recommending the use of nine. If three data points are used, the SEM around that point would envelop the projected growth rate of one word per week for an average third grader for 10 weeks. This becomes a factor if people make decisions about intervention alteration by comparing the location of data points to the aimline as is recommended in some field-based applications of the problem solving/RTI models (i.e., Heartland Area Education Agency, 2002). For example, if a decision rule is determined

that the intervention will be maintained unless 4 data points fall at or below the aimline, the overestimation of the baseline point could lead a team to change a successful intervention. Likewise, the underestimation of student baseline could lead teams to maintain an ineffective intervention.

In the current study, we reduced the variability of the probe set by including only those passages that met pre-determined cut-off scores. When only including probes that had an average score within +/-5 wcpm of the average, the SEM for a single passage was reduced from 18 to 12 wcpm for absolute decisions. When administering nine passages, the reduction in SEM was much smaller (i.e., reduced from 6 to 4 wcpm). The current data suggests that the most efficient way to reduce error may be to assess using sets of five probes that have been field tested and shown to deviate less than +/- 10 wcpm from the set average. Using a probe set constructed using these guidelines would allow practitioners to more efficiently estimate student performance by administering 5 probes (SEM of 6 wcpm) as compared to using 9 probes which have not been field tested and selected based on readability formulas (SEM of 6 wcpm).

Limitations and Future Directions

Although the current results have important applied implications, several limitations should be addressed. Current findings may lack generalization to practice for a variety of reasons. Given the number of probes administered, the sample size for the study was small. Thus variance due to passages may have been overestimated, while variance caused by student reading skill may have been underestimated. The population was homogenous with 100% of the subjects being Caucasian and from the rural Midwest. Future studies should include a large and more diverse group of participants.

Only third grade students and probes were examined in the current study using a single probe set of 20 passages downloaded from the dibels.uoregon.edu website (Good, Kaminski, & Dill, 2002). Students from other grades need to be studied to see if similar estimates of percentage of variance, reliability, and SEM are obtained. This particular probe set is generic in nature and was constructed with the goal of limiting the variability in passage difficulty through the use of readability formulas. The use of other probe sets would likely present different findings (Hintze & Christ, 2004). Also, CBM procedures specify the use of a median method to arrive at a score; however, the current study used averages to arrive at scores. Further research needs to be conducted using a median method with a variety of probe sets to investigate if there is a significant difference between the SEMs that accompany them.

It is not common practice to present students with five probes per day or 20 probes in one week. Furthermore, in the current study, the data was collected by one experimenter who presented five probes to students daily, at the same location, using uniform standardization procedures. This lack of variability in conditions may have decreased the amount of variability due to unaccounted sources of error. In applied settings, different people may assess different students (e.g., when collecting data for local norms) or the same student at different times (e.g., did the student respond to the intervention), which may increase error (e.g., Derr-Minneci & Shapiro, 1992). Probe administration for each student was randomized to equally disperse any unaccounted for main effects and/or interactions that made up the residual error estimate of 9%. Future researchers should conduct similar G theory studies using testing environment that are more typical in applied settings.

Conclusion

CBM can be used to provide data in a problem solving model to aid practitioners when making a variety of educational decisions ranging from which students should receive preventative interventions to whether a student is eligible for special education services (Deno, 1989). While the various uses of CBM in a problem solving model have ushered in a new wave of exciting possibilities, many new questions have arisen as well. As practitioners use CBM to address a variety of assessment questions, the psychometric characteristics for each purpose needs to be known, especially in light of its use to aid in eligibility determination. This includes both reliability estimates and a SEM that is aligned with whether relative or absolute decisions are being made.

The current investigation provides further evidence that wcpm is a reliable measure to produce a summative oral reading fluency score. However, the data also demonstrated that the SEM associated with wcpm may make it difficult to precisely measure and detect small differences between and within students. A relatively large source of error that can be controlled is passage variability within a probe set. The current data supports past research showing that reducing the variability of passage difficulty and/or increasing the number of probes given can lessen error and increase the confidence in the educational decisions made using wcpm data (Hintze & Christ, 2004). These studies converge to suggest that probe sets should be constructed with the goal of reducing passage variability and defining procedures about how many passages should be administered to make specific recommendations with the SEM associated with wcpm data in mind. Lastly, it is imperative that practitioners are aware of and have the

psychometric evidence available to understand and translate the error term for CBM given the purpose of assessment and if the decision being made is relative or absolute.

References

References

Allen, D. (1989). Periodic and annual reviews and decision to terminate special education services. In Shinn, M. R. (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 182-201). New York: Guilford Press.

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., Slider, N. J., & Williams, K. L. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review*, *33*, 218-233.

Bergan, J. R. (1977). *Behavioral consultation*. Columbus, OH: Charles E. Merrill.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

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

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*, 116-127.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Deno, S. L. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In Shinn, M. R. (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 1-17). New York: Guilford Press.

Deno, S. L., Fuchs, L. S., Marston, D. & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review*, *30*, 507-524.

Deno, S. L., & Mirkin, P. (1977). *Data-based program modification: A manual*. Reston, VA: Council for Exceptional Children.

Derr-Minneci, T. F. & Shapiro, E. S. (1992). Validating curriculum-based measurement in reading from a behavioral perspective. *School Psychology Quarterly*, *7*, 2-16.

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

Fuchs, L. S., & Deno, S. L. (1994). Must instructionally useful performance assessment be based in the curriculum? *Exceptional Children*, *61*, 15-24.

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

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

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

Good, R. H. & Kaminski, R. A. (2002). *DIBELS oral reading fluency passages for first through third grades* (Technical Report No. 10). Eugene, OR: University of Oregon.

Good, R. H., Kaminski, R. A., & Dill, S. (2002). DIBELS oral reading fluency. In R. H. Good & R. A. Kaminski (Eds.), *Dynamic Indicators of Basic Early Literacy Skills* (6th Ed). Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoregon.edu/.

Gresham, F. & Noell, G. H. (1998). Functional analysis assessment as a cornerstone for noncategorical special education. In D. J. Reschly, W. D. Tilly, & J. P. Grimes (Eds.), *Functional and Noncategorical Identification and Intervention in Special Education* (pp. 39-64). Des Moines, IA: Iowa Department of Education.

Gresham, F. M. & Witt, J. C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. *School Psychology Quarterly*, *12*, 249-267.

Hartmann, D. P., Roper, B. L., & Bradford, D. C. (1979). Some relationships between behavioral and traditional assessment. *Journal of Behavioral Assessment*, *1*, 3-21.

Heartland Area Education Agency. (2002). *Program manual for special education*. Johnston, IA: Author.

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

Hintze, J. M., Christ, T. J., & Keller, L. A. (2002). The generalizability of cbm survey-level mathematics assessments: Just how many samples do we need? *School Psychology Review*, *31*, 514-528.

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of g theory to curriculum-based measurement, *School Psychology Quarterly*, *15*, 52-68.

Howell, K. W., Fox, S. L., & Morehead, M. K. (1993). *Curriculum-based evaluation teaching and decision making* (2nd Ed.). Columbus, OH: Charles E. Merrill.

Kaminski, R. A. & Good, R. H. (1998). Assessing early literacy skills in a problem-solving model: Dynamic indicators of basic early literacy skills. In M. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 113-142). New York: Guilford Publications.

Kavale, K. A. & Forness, S. R. (1999). Effectiveness of special education. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of School Psychology* (pp. 984-1024). New York: John Wiley & Sons.

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

National Association of School Psychologists (2003). NASP recommendations:LD eligibility and identification for IDEA reauthorization. *Communique*, June, insert p. 2-6.

Powell-Smith, K. A., & Ball, P. L. 2002. Best practices in reintegration and special education exit decisions. In A. Thomas and J. Grimes (Eds.), *Best Practices in School Psychology IV* (pp. 1-20). National Association of School Psychologists: Washington, D.C.

Powell-Smith, K. A., & Bradley-Klug, K. L. (2001). Another look at the "c" in cbm: Does it really matter if curriculum-based measurement reading probes are "curriculum-based?" *Psychology in the Schools*, *38*, 299-312.

Reschly, D. J., Tilly, W. D., & Grimes, J. P. (Eds.). (1998). *Functional and noncategorical identification and intervention in special education*. Des Moines, IA: Iowa Department of Education.

Reschly, D. J., & Ysseldyke, J. E. (2002). Paradigm shift: The past is not the future. In A. Thomas and J. Grimes (Eds.), *Best Practices in School Psychology IV* (pp. 1-20). National Association of School Psychologists: Washington, D.C.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th Ed). San Diego, CA: Jerome M. Sattler.

Shapiro, E. S. (1996). *Academic skills problems: Direct assessment and intervention* (2nd Ed). New York: Guilford Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

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

Shinn, M. R. (Ed.). (1998). *Advanced applications of curriculum-based measurement*. New York: Guilford Press.

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

Shinn, M. R., Good, R. H., & Parker, C. (1998). Noncategorical special education services with students with severe achievement deficits. In Reschly, D. J., Tilly, W. D., & Grimes, J. P. (Eds.), *Functional and noncategorical identification and intervention in special education* (pp. 65-84). Des Moines, IA: Iowa Department of Education.

Shinn, M. R., Good, R. H., & Stein, S. (1989). Summarizing trend in student achievement: A comparison of methods. *School Psychology Review*, *18*, 356-370.

Shinn, M. R., Habedank, L., Rodden-Nord, K., & Knutson, N. (1993). Using curriculum-based measurement to identify potential candidates for reintegration into general education. *The Journal of Special Education*, *27*, 202-221.

Silva, F. (1993). *Psychometric foundations and behavioral assessment*. Newbury Park, CA: Sage.

Stage, S. A. (2001). Program evaluation using hierarchical linear modeling with curriculum-based measurement reading probes. *School Psychology Quarterly*, *16*, 91-112.

Tawney, J. W., & Gast, D. L. (1984). *Single subject research in special education*. Columbus, OH: Charles E. Merrill.

Tilly, W. D. & Grimes, J. (1998). Curriculum-based measurement: One vehicle for systemic educational reform. In M. Shinn (Ed.), *Advanced applications of curriculum-based measuremen*t (pp. 32-59). New York: Guilford Publications.

Tindal, G., Germann, G., & Deno, S. L. (1983). *Descriptive research on the pine county norms: A compilation of findings*. (Research Report No. 132). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Tindal, G., Marston, D., & Deno, S. L. (1983). *The reliability of direct and repeated measurement*. (Research Report No. 109). Minneapolis: University of Minnesota Institute for Research on Learning Disabilities.

Ysseldyke, J. & Marston, D. (1998). Origins of categorical special education services and a rationale for changing them. In D. J. Reschly, W. D. Tilly, & J. P. Grimes (Eds.), *Functional and Noncategorical Identification and Intervention in Special Education* (pp. 15-38). Des Moines, IA: Iowa Department of Education.

Appendix

Appendix A: Consent Forms

Parental Consent Form

Dear parent or guardian,

My name is Brian Poncy and I am a graduate student in the School Psychology Ph.D. program at the University of Tennessee. As part of my training, I am working with the children and teachers at Martensdale St. Marys Elementary School. I am requesting permission to involve your child in a study about Curriculum-Based Measurement, a test used to estimate a student's performance in reading. The Martensdale St Marys administration is fully aware of the project and has allowed me to conduct this research. The project involves having your child read a variety of grade level reading passages similar to the ones he or she is already reading in his or her classroom. It will take approximately 6-12 minutes per day for one week to complete this process.

I will not share your child's performance on these activities with any school personnel and his or her performance will not affect his or her school grades. The information gained by your child's participation may help develop more effective assessment strategies to aid school personnel in making decisions about student performance in reading.

I would greatly appreciate your permission to work with your child on this project. Please sign and date below if you would like your child to participate. Please fill in your child's name in the space provided and return the form to school. Thank you for your time and consideration.

If you any questions or concerns please contact me at brianponcy@yahoo.com or 515-554-9244.

Signature of Parent/Legal Guardian:_____ Date:_____

Child's Name:_____

Children's Assent Form

My name is Brian Poncy and I am a graduate student at the University of Tennessee. I am doing a project on reading. I will be working with other students at your school on this project. I would like for you to help me with this project. If you would like to help, I will need you to give me permission to include you. I will be spending about 5-10 minutes a day listening to you read from a variety of passages.

It is important for you to understand that your help is by choice. At any time, you can choose to no longer participate by informing your teacher or myself. If you have any questions please ask your teacher or me.

If you agree to help reading the stories, please mark the box next to "yes". If you do not want to help, then mark the box next to "no". Write your name on the line below. Thank you for your help.

Sincerely,

Brian Poncy

Yes

No

Name_____ Date _____

School Permission Form

Mr./Mrs. Principal,

My name is Brian Poncy and I am a graduate student in the School Psychology Ph.D. program at the University of Tennessee. I am requesting your permission to use the students and facilities at your school to conduct a study on the stability and accuracy of Curriculum-Based Measurement (CBM). The purpose of the study is to investigate the psychometric characteristics of assessment instrument and how it could be optimally used to for a variety of educational purposes and decisions (i.e., screening, defining student performance, eligibility).

I would need you to select two third grade classrooms to participate in the study. Once the classrooms were selected I would send permission forms to the parents to obtain their consent. In addition, before beginning the study I would explain the study to the students as they will have the option to decline to participate in the study. To complete the study I will need to see each of the participants once a day for five days for approximately 5-10 minutes per day. The time the student and I are in contact will be spent with the student reading a variety of grade level probes aloud. In addition, I will need permission to use the facilities at your school. Optimally this would include a small room that would normally be used for individualized testing but any available space with a limited amount of distractions could be used.

If you have any further questions or comments please contact me via email, brianponcy@yahoo.com, or by phone at 515-554-7288. I would greatly appreciate your permission to work in your school on this project. Please sign and date below if you would agree for your school to participate. Thank you for your time, effort, and cooperation.

Respectfully,

Brian Poncy

Signature of Building Principal

_____

Date: _____

Signature of District Superintendent

_____

Date: _____

Vita

Brian Poncy was born in Iowa City, Iowa on May 9, 1974. He was raised in Centerville, Iowa and graduated from Centerville High School in 1993. He attended the University of Northern Iowa, where he received a B.A. in Psychology (1998), M.A.E in Educational Psychology (1999), and an Ed.S. in School Psychology (2002). Brian is completing requirements for a Ph.D. in Education with a concentration in School Psychology. He currently works as an intern for Prairie Lakes Area Education Agency in Webster City, Iowa.