

MEASUREMENT ERROR IN PROGRESS MONITORING DATA: COMPARING
METHODS NECESSARY FOR HIGH-STAKES DECISIONS

A Dissertation

by

SUSAN ADELE DUPOISE BRUHL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2012

Major Subject: Educational Psychology

Measurement Error in Progress Monitoring Data: Comparing Methods Necessary for
High-Stakes Decisions

Copyright 2012 Susan Adele Dupoise Bruhl

MEASUREMENT ERROR IN PROGRESS MONITORING DATA: COMPARING
METHODS NECESSARY FOR HIGH-STAKES DECISIONS

A Dissertation

by

SUSAN ADELE DUPOISE BRUHL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approval by:

Committee Co-Chairs,	Patricia S. Lynch Kimberly J. Vannest
Committee Members,	Richard Parker James Kracht
Head of Department,	Victor Willson

May 2012

Major Subject: Educational Psychology

ABSTRACT

Measurement Error in Progress Monitoring Data: Comparing Methods Necessary for
High-Stakes Decisions. (May 2012)

Susan Adele Dupoise Bruhl, B.F.A., James Madison University;

M.Ed., University of Hawaii at Manoa

Co-Chairs of Advisory Committee: Dr. Patricia Lynch
Dr. Kimberly Vannest

Support for the use of progress monitoring results for high-stakes decisions is emerging in the literature, but few studies support the reliability of the measures for this level of decision-making. What little research exists is limited to oral reading fluency measures, and their reliability for progress monitoring (PM) is not supported. This dissertation explored methods rarely applied in the literature for summarizing and analyzing progress monitoring results for medium- to high-stakes decisions. The study was conducted using extant data from 92 “low performing” third graders who were progress monitored using mathematics concept and application measures. The results for the participants in this study identified 1) the number of weeks needed to reliably assess growth on the measure; 2) if slopes differed when results were analyzed with parametric or nonparametric analyses; 3) the reliability of growth; and 4) the extent to which the group did or did not meet parametric assumptions inherent in the ordinary least square regression model. The results indicate reliable growth from static scores can be obtained in as few as 10 weeks of progress monitoring. It was also found that within this dataset,

growth through parametric and nonparametric analyses was similar. These findings are limited to the dataset analyzed in this study but provide promising methods not widely known among practitioners and rarely applied in the PM literature.

DEDICATION

To my husband, Taborri, and my kids, Amelia, Martha, and Harrison

ACKNOWLEDGEMENTS

This dissertation is the culminating project at the end of a really long journey. I am thankful for where I've been and those who have been with me.

For my family, thank you for allowing me to carelessly venture down this road not knowing what might lie ahead. Personally, this project and pretty much everything else that has happened along the way has been mostly about what if I could do it, and what would I find?

Taborri, thanks for letting me follow my heart. For my children, I know you can achieve whatever you desire, and I hope you take journeys simply because you remain curious.

I am also humbled by what I've learned... and grateful for my committee---- Pat, Kimber, Rich and Jim, thank you for your wisdom, support, and sticking with me all this time.

TABLE OF CONTENTS

	Page
ABSTRACT	iii
DEDICATION	v
ACKNOWLEDGEMENTS	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES	ix
LIST OF TABLES	x
CHAPTER	
I INTRODUCTION.....	1
II A REVIEW OF THE LITERATURE: METHODS FOR ANALYZING AND SUMMARIZING PROGRESS MONITORING DATA	6
Introduction	6
Analytical Methods	12
Conclusion.....	19
III HIGH RELIABILITY AND PRECISION: USING STATIC PERFORMANCE TO ESTIMATE GROWTH FOR HIGH-STAKES DECISIONS.....	21
Introduction	21
Method	31
Results	36
Discussion	40
IV A COMPARISON OF MODELS FOR RELIABLY ESTIMATING GROWTH ON MATHEMATICS CONCEPTS AND APPLICATION MEASURES.....	44
Introduction	44
Method	53

CHAPTER	Page
Results	57
Discussion	64
V CONCLUSIONS	68
Implications for Practitioners.....	71
Implications for Progress Monitoring Systems.....	72
Implications for Future Research.....	73
REFERENCES	76
APPENDIX A	85
APPENDIX B	97
VITA	109

LIST OF FIGURES

FIGURE		Page
2.1	Calculation of the SEM_{eas}	10
2.2	Calculation of a confidence interval.....	11
2.3	Sample data set of mathematics problem solving probes over 15 weeks	13
2.4	Static performance methods at fixed points using (a) multiple trials or (b) approximate repeats.....	16
2.5	Confidence level of 84.3% around the Y_{est} score at week 3 and 12 with corresponding confidence interval overlap illustration.....	17
3.1	Time series linear regression over 20 weeks of progress monitoring	26
3.2	Static performance analysis of sample data set from 20 weeks of progress monitoring for repeated math concepts and application probes adjacent to weeks 3, 10, and 19 (a) and confidence limits at 83.4% around Y_{est} scores (b)	28
3.3	Confidence levels of 83.4% graphed around the Y_{est} at week 3, 10, and 19 with 95% CIs about the slope.....	30
4.1	A graph of one participant's scores across 20 weeks of PM that includes slope and R^2 index as calculated in a linear regression program	50

LIST OF TABLES

TABLE		Page
3.1	Participant Slope and Trendedness Indices on Third Grade M-CAPs	85
3.2	Summaries of Improvement Using Approximate Repeats Method	87
3.3	Sensitivity to Growth Using Approximate Repeats Method.....	91
3.4	Comparison of T-test Results to Overlapping Confidence Intervals Technique	92
4.1	Descriptive Results for Slopes	97
4.2	Individual Calculations of Slope Size, Trendedness and Standard Errors and Confidence Intervals.....	97
4.3	Slope Differences and Significance Levels.....	102
4.4	Tests of Residuals.....	107

CHAPTER I

INTRODUCTION

Progress monitoring (PM), repeated measurement of academic or behavioral competence over time, has gained widespread attention over the past decade as a result of the Response to Intervention model (RtI) (Foegen, Jiban, & Deno, 2007; Fuchs, 2004; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). PM is rooted in Applied Behavioral Analysis in the 1950s and Precision Teaching in the 1970s, but its application today is the result of studies of curriculum-based measurement (CBM) in the 1980s (Deno, 1985; Fuchs, 1986; Fuchs, 2004; Fuchs, Fuchs, & Stecker, 1989; Binder, 1990; Pennypacker, Gutierrez & Lindsley, 2003). Primarily PM has been recommended for monitoring student growth in the basic skills areas (i.e., reading, mathematics, writing, and spelling) to inform instructional decisions (Deno, 1985; Deno, 2003). More recently, scholars have suggested that PM results be used for high-stakes decisions (Deno, 2003; Foegen, et al., 2007; Fuchs & Fuchs, 1998; McGlinchey & Hixson, 2004; Restori, Gresham, & Cook, 2008; Wayman, et al., 2007), but questions are being posed about the capacity of PM and associated analyses to adequately summarize performance (Christ & Coolong-Chaffin, 2007; Parker, Vannest, Davis, & Clemens, 2010). With discord emerging in the PM literature over the use of PM for important decisions, this paper will provide an overview of the technical adequacy of CBM to assess growth. Included in the overview will be a specific focus on high-stakes decisions and the needed ingredients, identified

This dissertation follows the style of *The Journal of Special Education*.

by the literature, to appropriately summarize growth, setting the context for the present study.

Decades of research attest to the psychometric properties of PM measures. In 1985, Stan Deno (1985) proposed an alternative to commercially available standardized achievement tests-- curriculum-based measures (CBM). Deno argued, that unlike achievement tests that were disassociated from the curricula and instructional decision-making, CBM had a number of salient features for monitoring student progress toward year-end goals. He described how he and colleagues designed CBM to be reliable and valid indicators of student achievement, simple and efficient to use, easily interpreted and inexpensive. For decades, these features have contributed to the viability of progress monitoring measures and their capacity to support low-stakes decision-making (e.g., informing instructional decisions), with respect to children with disabilities. When reform efforts intensified in the late 1990s and early 2000s, scholars' recommendations shifted. Based on theory, research, and policy, scholars are recommending PM be used to increase services or make program placement decisions within a framework known as the Response to Intervention (RtI) model (McGlinchey & Hixson, 2004; Restori, Gresham, & Cook, 2008).

The challenge of PM within the context of RtI is that it is being recommended for important decisions, but current research indicates PM results are less reliable and precise than previously reported (Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008; Hintze & Christ, 2004; Hintze, Owen, Shaprio, & Daly, 2000; Jenkins, Graff, & Miglioretti, 2009; Poncy, Skinner, & Axtel, 2005). The difficulty with using PM for high

stakes decisions is that the measures are highly sensitive to small changes in performance over time, and this impacts the reliability and consistency of the results (Christ & Coolong-Chaffin, 2007). The basis for reliability reporting is Classic Test Theory (CTT). CTT is sufficient, per the research, for low-stakes decisions e.g. the equivalency of PM measures for screening and in-class decisions (Deno, 1985; Deno, 2003). When PM results are used to guide high-stakes decisions (e.g. placement in supplemental instruction or special education) CTT-based measurement reliability is inadequate for repeated measurement. Also results for high-stakes decisions typically include an expression of error and a level of confidence needed to express the likelihood of a “regretted decision” (Ardoin & Christ, 2009; Christ & Coolong-Chaffin, 2007; Hintze & Christ, 2004). The current study will apply alternative approaches to obtaining more reliable and precise results for important decisions.

The dissertation is organized into five chapters including this introduction. Chapters II through IV are intended to be stand-alone manuscripts acceptable for publication. This dissertation will first present the context of progress monitoring from its roots to its current uses in the field. Chapter II is a literature review summarizing the theoretical underpinnings, research, and policy driving progress monitoring to inform important decisions. This chapter explores the technical adequacy of PM measures based on Classic Test Theory and issues with this model when measures are used repeatedly. Other models are presented that are more suitable for reporting the reliability of PM results. Chapters III and IV present a two-part study that specifically focuses methods that produce more sound analysis and summaries of progress monitoring (PM) results

for high-stakes decisions. The study was conducted using a dataset of “low performing” third graders’ Mathematics Concepts and Applications (M-CAP) PM results obtained from AIMSweb; a nationally recognized and respected repository of PM measures and student results. Part one of the study, presented in Chapter III, examines how much time is necessary to reliably measure change among the “low performing” participants. A visual analysis technique known as “approximate repeats” was used in the study to measure growth between two point estimates obtained from an ordinary least squares (OLS) model. In Chapter IV, using the same group of participants, the second portion of the study compares parametric analytic techniques, the commonly applied ordinary least squares, to nonparametric analyses, Theil-Sen and Tau (Conover, 1980; Hollander & Wolfe, 1999; Sprent, 1993) to determine if slopes were similar and the extent to which the summaries of performance are reliable. The study contained here seeks to answer several pressing questions with respect to using PM for important decisions:

- 1) Within an OLS regression model, what span of time is required to measure reliable change in performance (static measure of progress) from Y_{est} scores and their confidence intervals.
- 2) Among low performing students on the M-CAP, how closely do slopes when calculated with the Theil-Sen method approximate OLS regression slopes?
- 3) How reliable are the slopes as indicated by trendedness indices, standard errors and ratios of score/SEslope among low performing students on the M-CAP?
- 4) To what extent do dataserries among low performers fail to meet parametric assumptions of normality and equal variance?

Based on the outcomes of this study the final chapter, Chapter V, will examine ways in which answers to the research questions may contribute to defensible methods for informing high-stakes decisions. Implications for future research and limitations of the current study will be described. Also applications for practitioners in the field will be shared.

CHAPTER II
A REVIEW OF THE LITERATURE: METHODS FOR ANALYZING AND
SUMMARIZING PROGRESS MONITORING DATA

Introduction

Progress monitoring (PM) refers to a set of assessments using curriculum-based measures to monitor student growth and inform decisions (Deno, 1985). In the last decade PM has shifted considerably as a result of policy and research agendas dedicated to increased accountability and improved student outcomes (Fuchs, 2004; IDEIA, 2004; U.S Department of Education, 2002). PM has become the vehicle for ensuring students are not only making progress on annual goals but also improving outcomes on high-stakes assessments (McGlinchey & Hixson, 2004; NCLB, 2001; Silberglitt & Hintze, 2005). School teams are now using PM results to make decisions about increasing services and placement (Case, Speece & Molloy, 2003; Fuchs, Fuchs, Compton, Bryant, Hamlett & Seethaler, 2007; Speece & Case, 2001). Further, PM for the purposes of eligibility is intensifying in policy debates and practice due to questions about the efficacy of the discrepancy model (Carnine, 2003; Fuchs & Fuchs, 1998; IDEIA, 2004).

In 1998, Fuchs & Fuchs presented a compelling alternative to the discrepancy model that thrust the decision-making power of PM to new levels. The traditional discrepancy approach relies on a student's cognitive and achievement test results at a static point in time relative to typically achieving peers. Critics of the discrepancy model

argue that cognitive and achievement tests are lengthy, are not predictive of academic achievement, and lack the capacity to inform instructional decision-making or monitor growth toward year-end goals (Fuchs & Fuchs, 1998; McGlinchey & Hixson, 2004; Restori, Gresham, & Cook, 2008). The alternative, the Response to Intervention (RtI) model, relies on curriculum-based measures (CBMs) for PM that assesses an individual's static score at a point in time and improvement rate over time. CBMs are quick to administer and are useful in terms of determining if a student is responding to changes in instruction. Students who remain unresponsive in this model are identified for more intensive interventions including those provided in special education. The integrity of this model relies, however, upon the technical adequacy of the PM measures.

The technical adequacy of PM measures is well documented and has been relatively unchanged up until the last decade (Foegen, Jiban & Deno, 2007; Wayman, Wallace, Wiley, Ticha & Espin, 2007). Recently some scholars have argued that PM measures are not reliable enough to inform high-stakes decisions (Ardoin & Christ, 2009; Christ & Coolong-Chaffin, 2007; Hintze & Christ, 2004). The issue is that technical adequacy studies have relied on Classic Test Theory (CTT) as the basis for determining if measures are reliable (Deno, Fuchs, Marston, & Shin, 2001; Marston, 1989). CTT-based reliability is appropriate when measures are administered one or two times (e.g., benchmark assessing in the fall and winter) and low-stakes decisions (e.g., in-class instructional changes). CTT is not appropriate for repeated measurement (i.e., progress monitoring) especially in instances where medium- to high-stakes decisions are being considered (Deno, et al., 2001; Nunnally, 1967; Salvia & Ysseldyke, 2003). The

remainder of this paper will explain why and describe other models more suitable for PM.

Classic Test Theory

PM procedures, developed more than three decades ago, apply CTT-based reliabilities to ensure that the measures are technically sound. CTT, also known as true score theory, is a statistical model for comparing test scores (e.g., alternate form or retest) and determining the extent to which the results can reliably estimate performance and minimize error (Hambleton & Jones, 1993). The true score is a hypothetical mean score that would occur if one took a test an infinite number of times (Nunnally, 1967). Measurement error (e.g., random effects of test fatigue or distractions during the test) is documented in CTT through such procedures as retest or alternate form reliability. Retest reliability is obtained by administering a single measure to a group of subjects at two different points in time, not less than two weeks apart according to most experts. Alternate form reliability is achieved by creating equivalent measures and administering them to a group of subjects. Progress monitoring research has primarily relied upon CTT as the basis of technical adequacy studies. CTT-based reliability indices from alternate form or retest are most commonly reported in PM studies as noted in two recent meta-analyses in the literature (Foegen, et al., 2007; Wayman, et al., 2007). Foegen, et al. (2007) identified 32 studies of mathematics progress monitoring measures. Twenty-two studies included technical adequacy results and all but two used traditional CTT-based reliability approaches. The coefficients ranged from 0.48 to 0.99 with most scores being considered moderate to highly reliable. Similarly, Wayman, et al. (2007) reviewed 66

studies of PM in reading of which 16 studies reported reliability results. Ten of the sixteen studies applied alternate form or test-retest reliabilities and the coefficients ranged from .56 to .99, again being considered moderate to highly reliable. Summaries and interpretations under CTT have limits that impact the level of decisions that can be made, but these limits have not been addressed in the literature until recently.

CTT-based reliabilities, as reported in technical adequacy studies, are determined by calculating group standard deviations and correlating the score sets (e.g., comparison of results of a measure between two groups of students). Reliability in CTT is useful for summarizing the expected error between one or two test administrations (Nunnally, 1967; Salvia & Ysseldyke, 2003). Reliability indices from CTT provide an indication of test quality and score precision for the purposes of screening or modifying the instruction.

Score precision is expressed as the standard error of measurement (SEM_{eas}) because it is the standard deviation of the sample distribution. The SEM_{eas} is used to estimate error around an obtained score as shown in Figure 2.1 (Nunnally, 1967; Thompson, 2006). A tenet of CTT is that an individual's obtained score is subject to variation if, in theory, a test is administered a number of times. The SEM_{eas} is calculated as, $\sigma_{\text{meas}} = \sigma_x \cdot \sqrt{1 - r_{xx}}$ where r_{xx} is a reliability index from alternate form, split-half, or retest procedures (Nunnally, 1967). The SEM_{eas} is used to determine if there is a significant difference between scores (i.e., an individual score and the mean of all other scores, two individual scores, or an individual's scores administered at two different times). Figure 2.1, modeled after a description in Nunnally (1967), demonstrates how the

SEM_{eas} is used to calculate an estimation of the standard deviation around an individual's obtained score. The SD is an index of the amount of error, meaning a smaller SD indicates fewer errors. Fewer errors mean more reliability and an indicator of the extent to which measures are repeatable e.g, use of equivalent forms for PM. For example, an individual obtains a score of 20 on a math CBM. The test manual indicates the alternate form reliability is $r = .81$, the sample mean was 24 items correct, and the standard deviation of scores (the sum of all variations from each test mean) is 10. The SEM_{eas} provides a measure of the population variance as shown in Figure 2.1.

$$\begin{aligned}
 \sigma_{\text{meas}} &= 10 * \text{sqrt} (1 - .81) \\
 &= 10 * \text{sqrt} (.19) \\
 &= 10 (.436) \\
 &= 4.36
 \end{aligned}$$

Figure 2.1. Calculation of the SEM_{eas} .

High stakes tests usually express error with a typically set CTT-based confidence level at 90-95%. The SEM_{eas} is then used to establish a confidence band in which the individual's true score is likely to reside. Confidence Intervals (CIs) can be obtained two ways: by calculating the confidence interval symmetrically about the obtained score or by using the SEM_{eas} to obtain an unbiased estimate of the true score (Nunnally, 1967). Both are considered acceptable but the simplest is to use the SEM_{eas} about the obtained

score to establish confidence levels as shown in Figure 2.2. In this example, with the typically selected confidence level of 95% represented by the z value 1.96, the confidence interval is calculated symmetrically about the individual's obtained score. The confidence level is interpreted as, one can be 95% confident that the true score is likely to be between 11.06 and 28.94 (see Figure 2.2).

$$20 \pm 4.56 * 1.96 = 11.06 \text{ to } 28.94$$

Figure 2.2. Calculation of a confidence interval.

In the example detailed in Figures 2.1 and 2.2, CTT-based reliability and associated confidence intervals are useful in terms of making judgments about test quality, score precision, and change between scores among individuals across one or two test administrations e.g., screening students in the fall and winter (Nunnally, 1967; Salvia & Ysseldyke, 2003). Practitioners should be cautioned that reliance on a CTT SEM_{eas} for PM is insufficient. The SEM_{eas} from alternate form or retest reliabilities are not appropriate for more than one or two test administrations, and it does not account for additional sources of error inherent in repeated measures over time.

Progress monitoring measures are sensitive to additional sources of error such as variations in individual performance over time and probe difficulty that impact reliability (Deno, et al., 2001; Poncy, Skinner, & Axtell, 2005). A small number of studies, almost exclusively limited to oral reading fluency PM measures, have examined sources of error present in PM that impact results including: variations in testing location and staff (Derr

& Shapiro, 1989; Derr-Minneci & Shapiro, 1992); developmental differences (Hintze, Daly & Shapiro, 1998; Poncy, et al., 2005); passage difficulty (Francis, Santi, Barr, Fletcher, Varisco & Foorman, 2008; Hintze, Daly, & Shapiro, 1998; Hintze, Owen, Shapiro & Daly, 2000) and the frequency of monitoring progress (Hintze, et al., 1998; Hintze & Christ, 2004; Jenkins, Graff & Miglioretti, 2009). Measurement variation is compounded across multiple administrations, when error estimates are limited to CTT-based SEM_{eas} , causing inflated performance results and faulty decision-making (Hintze & Christ, 2004). Further, the studies identified above question the adequacy of CTT reliability estimates demonstrating that for PM applications, most collections of equivalent probes provide lower reliability and lower score precision than previously estimated under CTT. These studies have seriously questioned the adequacy of CTT reliability estimates for PM applications. Increasingly, researchers are recognizing the need for better estimates from outside the CTT model (Hintze et al., 1998; 2000; Parker, Vannest, Davis & Clemens, 2010; Poncy et al., 2005).

Analytical Methods

Time Series Linear Regression

Time series linear regression (TSLR) is one model that provides a more accurate and precise estimate of PM performance (Christ, 2006; Good & Shinn, 1990; Parker, et al., 2010; Shin, Good & Stein, 1989). TSLR is a regression model that describes a statistical relationship between variables in which the variable of time is a constant that can serve as a predictor variable (Neter, Kutner, Nachtsheim, & Wasserman, 1996). The strength of the relationship between the variables in TSLR is through the estimate of

least squares which uses calculus to minimize the squared distance of scores from a regression line.

The TSLR model has several advantages for analyzing growth. The results are predictive of future performance and used extensively in business, industry, and the social sciences (Shinn, et al., 1989; Neter, et al., 1996; Draper & Smith, 1996). This model is also familiar to practitioners and easily accessible because of commercially available PM systems such as AIMSweb, interventioncentral.org, and DIBELS. Because of these PM systems, practitioners should be familiar with examining graphs of time series data (as in Figure 2.3) to analyze and summarize PM performance.

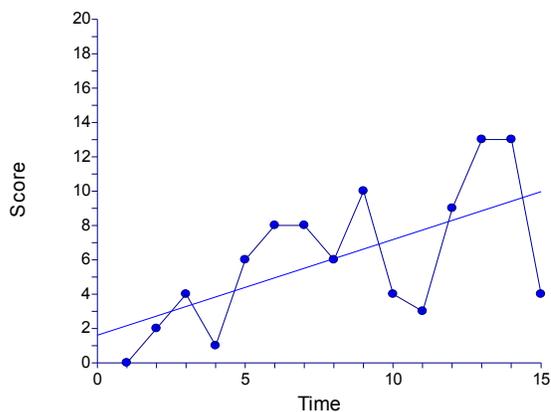


Figure 2.3. Sample data set of mathematics problem solving probes over 15 weeks.

Another advantage of TSLR is that the model can reliably estimate static performance and improvement rate over time. The best and most accurate estimate of performance is a point on the regression line known as the Y_{est} or Y_{hat} (Christ, 2006; Neter et al., 1996). The Y_{est} is an estimated score of performance that is neither the

obtained score at a particular week nor an estimated score influenced by other scores within close proximity. The Y_{est} (or \hat{Y}) is an estimated mean score (a parameter) at any particular value of X which is time expressed as $\hat{Y} = b_0 + b_1X$ (Draper & Smith, 1998). Error in TSLR can come from multiple sources, unlike in CTT where the SEM_{eas} is calculated from a group SD and a reliability index obtained from two test administrations. The standard error of the Y_{est} is calculated from the regression line and accounts for additional sources of error not accounted for by the CTT model (Christ & Coolong-Chaffin, 2007; Draper & Smith, 1966; Franklin, Allison & Gorman, 1997; Neter, et al., 1996; Parker et al., 2010). Also, the TSLR model is more flexible, as it can select as a Y_{est} score any point on the regression line and can calculate the standard error (SE) for any of those points. The Y_{est} score, and its SE are also more robust, as its calculation benefits from including all data points in the time series. Finally, the Y_{est} and its SE are based on an individual's performance rather than the performance of a particular peer group, cohort, or norm-group for interpretation (through the group SD). This heavy reliance on a group SD makes CTT SEM_{eas} unstable from one study to the next (Thompson, 2007). The Y_{est} provides a reliable and accurate estimate of static performance or growth over time necessary for medium- to high-stakes decisions (Hintze & Christ, 2004).

Additional Methods within the TSLR Model

The TSLR model has limits including that adjacent scores can be highly correlated and is very sensitive to outliers (Shinn, et al., 1989). Also for individual progress monitoring, the Y_{est} scores are mean estimations, and the favorable SEs for those scores is therefore

constructed for means. In individual PM results, the scores are not means, but individual data points from the individual client. There are two solutions to this problem of calculating a favorable SEs and useful CIs around Y_{est} scores, both from applied regression textbooks from outside the social sciences (Draper & Smith, 1998; Neter et al., 1996). The first solution is to obtain results from “multiple trials” or probes. Referring to Figure 2.4A in this instance three probes were administered at week 3. The best estimate at week 3 is not the mean or median of the scores. The Y_{est} of 2.61 is the best estimate of performance at week 3 and is calculated with a regression program which provides this output. Using the multiple trials method, the three scores are treated as three weights at time X, along with all other single data points at their individual X times. The second solution is a method termed by Draper & Smith (1998) as “approximate repeats” (shown in Figure 2.4B). The “approximate repeats” method may be more feasible from a practitioner’s stand point. In this method one selects adjacent scores e.g. scores from weeks 13, 14 and 15 and enters all scores at week 14. The regression program calculations treat these scores as weights, as in the “multiple trials” approach. From those scores a Y_{est} is derived and 8.55 items correct serves as the best estimate of performance at week 14. A disadvantage of the “approximate repeats” method is the loss of two data points at other X times when they are consolidated to a single middle X time.

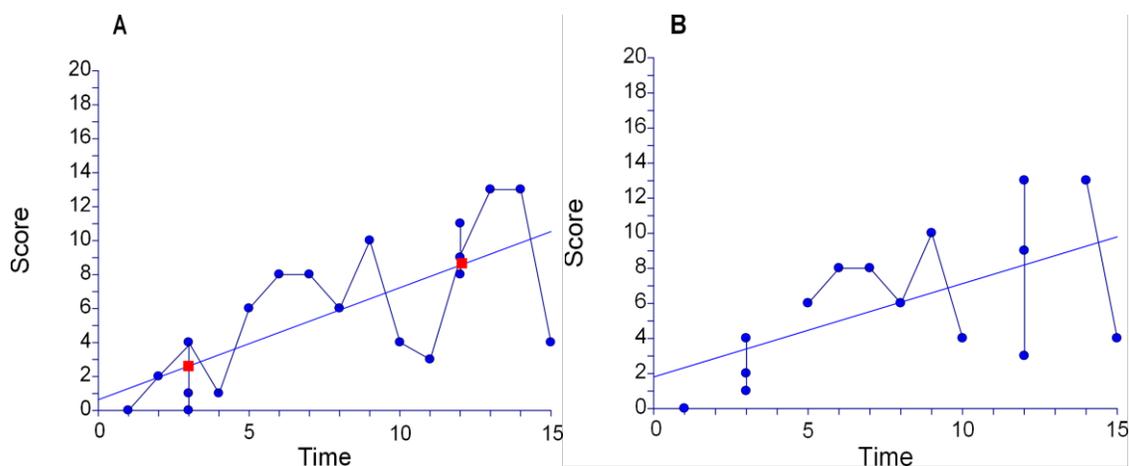


Figure 2.4. Static performance methods at fixed points using (a) multiple trials or (b) approximate repeats.

The calculation of Y_{est} scores and their SEs permits comparison of growth between two static points in time (Y_{est} scores at two X values). It also permits calculation of differences between two Y_{est} scores from two different data streams, such as two different students when progress is being monitored. Finally, it permits calculation of the difference between an individual's Y_{est} score and a group mean score; for example, comparing an individual's Y_{est} score at a particular point in time with a group score from a one-shot administration of the measure.

Another issue with PM summaries is that they are typically limited to summaries of slope (Foegen et al., 2007; Wayman, et al., 2007). For medium to high stakes decisions, slope should be accompanied by measurement error, which is the SE for the slope or SEb (Christ & Coolong-Chaffin, 2007; Parker et al., 2010). In Figure 2.3 the individual's obtained rate of improvement is evident in the trend line, but given the variability in results what degree of certainty does one have that the improvement is real or simply due to chance? The reliability of improvement rate can be calculated two

ways. First reliable growth can be reported using the standard error of slope (SEb) which is “the line of best fit” expected as a result of the standard error of estimate (Hintze & Christ, 2004). SEb is used to calculate both p-values for the slope corresponding

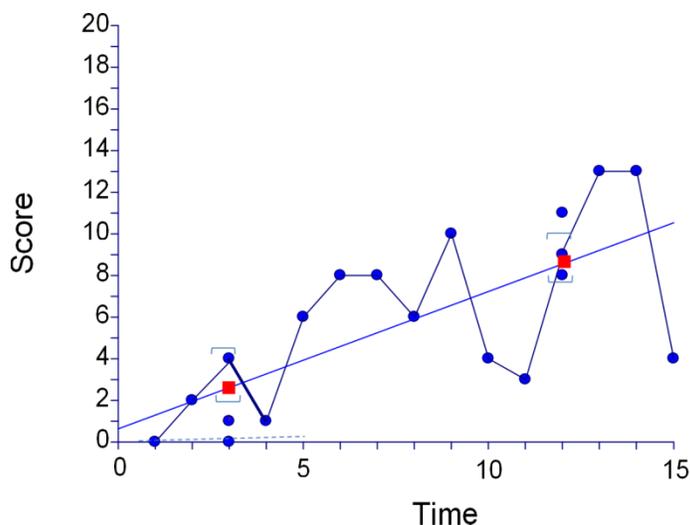


Figure 2.5. Confidence level of 84% around the Yest score at week 3 and 12 with confidence interval overlap illustration.

and confidence intervals around the slope. Confidence intervals can also serve as a quick-and-easy significance test, because if the CIs do not include zero then the slope is significantly different from zero (at the chosen confidence level). SEb is calculated using the following function: $s^2\{b_1\} = \text{MSE} / \sum(X_i - \bar{X})^2$. These procedures minimize the discrepancies between observed and predicted values. The second way to reliably report growth is through the examination of overlapping confidence intervals as shown in Figure 2.5 (Schenker & Gentleman, 2001). In figure 2.5 a static performance level at

weeks 3 and 12 was estimated from the time series regression model. The model also reports error and confidence intervals that were set at 83.4% and plotted on the graph. Significant growth can easily be obtained by visually inspecting the presence of non-overlapping confidence intervals as shown in Figure 2.5. When PM results are used to guide more serious educational decisions, placement in supplemental instruction or special education, measurement precision and an expression of the likelihood of a “regretted decision” are necessary (Ardoin & Christ, 2009; Christ & Coolong-Chaffin, 2007; Hintze & Christ, 2004).

Confidence intervals for an individual Y_{est} or slope can be set for any number of confidence levels to match the decision-making context. One’s willingness to make a regretted decision would be small in a situation that has social and financial consequences (Thompson, 2006; Parker, et al. 2010). For example, placement in special education has social, financial, and legal consequences, so setting a confidence level at 95% would indicate that the willingness to accept a regretted decision in 5% or fewer instances. A lower level of confidence may be acceptable for some PM decisions. Lower confidence levels of 80%, 85%, or 90% are a better match for in-class or reversible decisions.

Application of Nonparametric Analysis

A nonparametric alternative to the OLS model, applied in other fields such as business, health, and the earth science, where trends are not always linear, is Theil-Sen and its trendedness index, Kendall’s Tau (Conover, 1980; Sprent, 1993). Unlike OLS which assumes a normal distribution, Tau belongs to a family of distribution free

analyses (Sprent, 1993). Theil-Sen, analogous to linear regression slope, is a calculation of the median slope of all the data pairs within the set. Unlike linear regression in which the line of best fit is an estimated mean of the data pairs over time, Theil-Sen is a line that fits through the median of pairs of data (Conover, 1980; Sprent, 1993). Theil-Sen slope is calculated by finding all possible slopes where N is all data points and expressed $[N*(N-1)]/2$. Slopes for all data pairs are then calculated using the algorithm for slope $Y_b - Y_a / X_b - X_a$, the scores for data points “a” and “b” for all X and Y data pairs. The median value of all the mini slopes, the Theil-Sen slope, is given by $a_i = y_i - b*x_i$ (Sprent, 1993). Kendall’s Tau is a nonparametric trendedness index, analogous to R^2 , and provides a rank correlation coefficient equal to 1 when all data pairs show improvement. Tau provides an index of stability for PM data in terms of providing a percentage of data that show improvement (Conover, 1980; Parker et al., 2010; Sprent, 1993). It is calculated as follows: $[\# \text{ of pairs that show improvement} - \# \text{ of pairs not showing improvement} / \text{total number of pairs}]$. This nonparametric alternative is useful in PM where significant score variation or bounce and outliers are common (Weiss, 2003).

Conclusion

In sum, measurement error estimated from CTT is limited in a three distinct ways. CTT-based reliability: 1) does not account for variations in student performance over time, 2) is obtained from one or two test administrations, and 3) has an unstable SEM_{eas} calculated from a group SD and reliability score. Other methods for analyzing and summarizing PM results are more appropriate for informing high-stakes decisions. Such methods as TSLR calculate level and trend more precisely by accounting for error

across multiple data points. The model also provides more reliable performance estimates at a point in time through the Y_{est} and over time through a slope estimate (SEb). Precision is also enhanced when one can report error and set confidence limits fitting the educational decision (i.e., 95% CIs are often used for important decisions). Non-parametric models offer options other than OLS for summarizing PM results that may be non-linear. Given the range of decisions made with PM data from mere instructional modifications (low-stakes) to changes in placement that have social, financial and legal consequences (high-stakes), practitioners need to be knowledgeable of the limits of CTT-based reliabilities and skills to implement other analytic techniques.

CHAPTER III
HIGH RELIABILITY AND PRECISION: USING STATIC PERFORMANCE TO
ESTIMATE GROWTH FOR HIGH-STAKES DECISIONS

Introduction

Progress monitoring (PM) results from curriculum-based measurement (CBM) are being considered in high-stakes decisions, but scant research has attended to the reliability and precision of PM results for that purpose. Instead the suitability of CBM for PM has focused on two stages of CBM research over several decades (Fuchs, 2004). Both stages are devoted to identifying students in need of early intervention through periodic screening of performance at a point in time (Stage 1) and growth over time (Stage 2). Fuchs states, “Recent CBM research may focus disproportionately on Stage 1,” (p. 191), but she stresses that the traditional psychometric properties have shown the measures to be reliable and valid (Fuchs, 2004). According to the literature, less scholarly attention has been devoted to Stage 2 research (Fuchs, 2004; Foegen, Jiban, & Deno, 2007; Wayman, Wallace, Wiley, Ticha, & Espin, 2007), and an emerging body of research is questioning the capacity of CBM to reliably monitor progress over time (Derr & Shaprio, 1989; Derr-Minneci & Shapiro, 1992; Dunn & Eckert, 2002; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008; Hintze & Christ, 2004; Hintze, Daly, Shaprio, 1998; Hintze, Owen, Shapiro, & Daly, 2000; Jenkins, Graft, & Miglioretti, 2009; Poncy, Skinner, & Axtell, 2005). Others are suggesting addition methods for analyzing and summarizing PM results (Christ & Coolong-Chaffin, 2007; Deno, 2003;

Parker, Vannest, Davis, & Clemens, 2010). Methods outside the field of education offer approaches for using static performance levels to estimate growth (Draper & Smith, 1998; Payton, Greenstone, & Schenker, 2003; Schenker & Gentleman, 1999) that may be informative for both stages of PM research.

This paper presents part one of a two-part study examining methods for summarizing PM data to reliably and precisely inform high-stakes decisions. An overview of the PM literature indicates how reliability is typically obtained. Reliability under progress monitoring conditions for the purpose of high-stakes decision-making must be obtained differently. Some alternate methods, many of which are new to education, will be presented in this introduction. In the methods and results section these new methods were applied to an extant PM dataset to determine how many weeks were needed to reliably assess progress using static performance levels.

Traditional Methods for Reliably Reporting Growth

The reliability of PM to measure static performance, the focus of Stage 1 research, has relied upon Classic Test Theory (CTT) as shown in two recent meta-analyses (Foegen, et al., 2007; Wayman, et al., 2007). Reliability in CTT is determined using such methods as retest or alternate form reliabilities to assess if scores can be replicated across administrations. A recent study by Clark and Shinn (2004) provides an example of how reliability under the CTT model has been applied to CBM and ultimately PM studies. Clark and Shinn's study of the development of early numeracy CBM reports high rates of alternate form reliabilities ranging from 0.78 to 0.90 and retest reliabilities yielding coefficients of approximately 0.80 at weeks 13 and 26. This

study noted, consistent with texts on psychometric theory, that reliabilities were sufficient for the purposes of screening and decision-making (Clark & Shinn, 2004; Nunnally, 1967). In CTT, estimations of error, or the SEM_{cas} , are calculated from a group standard deviation and a reliability index from one or two test administrations. Fewer errors mean more reliability and indicate the extent to which measures are repeatable (e.g., use of equivalent forms for PM) (Nunnally, 1967). CTT is routinely applied to high-stakes assessments, such as an IQ test, where a test may be administered one or two times (Nunnally, 1967).

Issues of Reliability in Progress Monitoring

CTT-based reliability is inadequate for PM according to a small body of emerging literature in the area of oral reading fluency. When measures are administered repeatedly, variables such as developmental differences or passage difficulty can vary, and in some studies the results show reliability indices that are lower and less precise (Derr & Shapiro, 1989; Derr-Minneci & Shapiro, 1992; Dunn & Eckert, 2002; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008; Hintze & Christ, 2004; Hintze, Daly, & Shapiro, 1998; Hintze, Owen, Shapiro & Daly, 2000; Jenkins, Graft, & Migloretti, 2009; Poncy, Skinner & Axtell, 2005; Schatschneider, Wagner, & Crawford, 2008). Specifically, Hintze, et al. (1998) found that developing readers, when given a difficult passage, demonstrated more fluency problems than when the passage more closely matched their performance level. Francis, et al., (2008) raised concerns that when measures are not equivalent, gains or losses are misinterpreted. In their examination of the equivalency of oral reading fluency measures, the researchers found that passage

difficulty impacted individual results by 26 words read per minute (Francis, et al., 2008). Schatschneider et al. (2008) administered orally reading fluency probes in four sessions between the months of September and April. In each session the first grade participants were administered three oral reading fluency probes and the median score was used as the final data point for that testing session. The study found that improvement rates were less reliable and not predictive of year-end reading performance when compared to other year-end measures. The authors posed that more administrations over the year may have yielded higher reliability coefficients (Schatschneider et al., 2008). More data points and controlled conditions did increase reliability in two studies by Hintze, et al., (2000). Hintze et al. (2000) administered oral reading fluency probes twice weekly for eight weeks in the first study and twice weekly for ten weeks in the second study. The second study also varied passage difficulty, alternating administrations with probes on grade level and probes on goal level. Again outcomes were impacted by passage difficulty, but when such variables were controlled, reliable estimates of performance were found in as few as eight to ten data points. These studies indicate that CTT-based reliabilities reported in the literature are appropriate for interpreting reliability between one or two test administrations but insufficient for repeated measurement.

Time Series Linear Regression (TSLR)

A model suitable for reliably estimating growth is the ordinary least squares regression (OLS) model, also known as time series linear regression (TSLR) when the regression analysis includes time series data as with PM results (Neter, Kutner, Nachtsheim, & Wasserman, 1996). The ordinary least squares (OLS) model has been

applied to Stage 2 research to address the technical features of slope (Deno, Fuchs, Marston, & Shin, 2001; Foegen et al., 2007; Fuchs, 2004; Good & Shinn, 1990; Shinn, Good & Stein, 1989; Wayman et al., 2007). Linear regression is ideal for analyzing the variability within repeated measurement over time. Sources of error such as those found in the oral reading fluency studies are not identified individually in this model, but Parker, et al. (2010) assert that this is unnecessary as long as measurement error is known.

Practitioners should be accustomed to examining student growth using several available computerized software systems such as interventioncentral.org, AIMSweb, and DIBELS, but little guidance from the field supports practitioners much beyond visual interpretations of PM results. For example, PM systems graph growth rates or slopes for individual students that are useful for goal setting and monitoring progress toward goals as shown in Figure 3.1. This figure graphically displays a student's growth over 20 weeks of progress monitoring. While slope has been considered the primary parameter of growth (Fuchs & Fuchs 1998; Fuchs, Fuchs, Hamlett, Waltz, & Germann, 1993), practitioners are limited to comparing the student's growth rate to normed peer results. The slope index has several limitations for reliably and precisely estimating either static performance or rate of improvement over time. Other indices from the OLS model provide more reliable information about the stability of the results, performance level and trend.

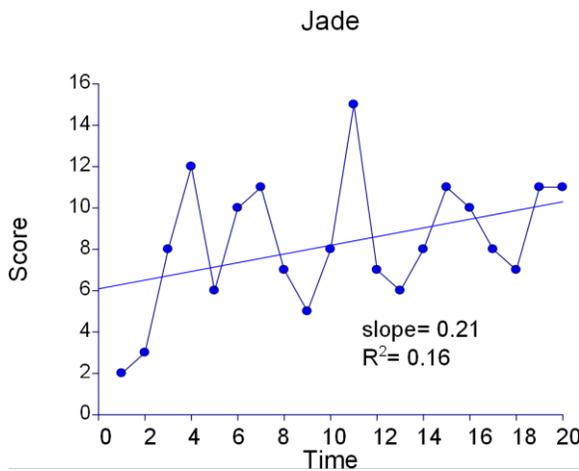


Figure 3.1. Time series linear regression over 20 weeks of progress monitoring.

Issues of Stability

A small body of literature and texts on linear regression analysis provide key data needed to account for score variability and summarize performance level (Christ, 2006; Draper & Smith, 1998; Kazdin, 1982; Neter et al., 1996; Payton et al., 2003; Parker et al., 2010; Schenker & Gentleman, 2001). Score variation, as seen in Figure 3.1 reflects a lot of “bounce” in the data making it difficult for practitioners to estimate and adequately summarize performance (Shinn, Good, & Stein, 1989). Trendedness indices are useful in explaining score variability and describing the extent to which the student’s scores follow the trend line or the consistency in which scores contribute to growth over time (Christ & Coolong-Chaffin, 2007; Parker et al., 2010). If one were to summarize results shown in Figure 3.1, the slope is 0.21 items correct per week. Normed peer results suggest this is typical growth rate for this PM measure but the R^2 index is low, indicating only a small percent (16%) of the student’s scores are contributing to growth.

Summarizing Performance Level

Score variability also makes it difficult to predict performance at a point in time. Typically estimations of the current level of performance (static score) are based on average or median scores (Christ & Coolong-Chaffin, 2007; Fuchs & Fuchs, 1998; Marston, 1989). Medians or averages are fine for low-stakes decisions, but medium- to high-stakes decisions should include reliable and precise estimates of performance that include error and express confidence (Christ & Coolong-Chaffin, 2007; Parker et al., 2010). In the OLS model, a point estimate known as the Y_{est} can be used instead of a median score or slope to summarize performance at a point in time and over time (Draper & Smith, 1996; Neter et al., 1996; Parker et al., 2010). The Y_{est} or \hat{Y} is a point on the regression line that is the best and most accurate estimate of performance because it is an estimated mean score, based on all of the scores in the sample, at a particular value of X (time) and expressed as $\hat{Y} = b_0 + b_1X$ (Draper & Smith, 1998; Neter et al., 1996). The standard error (SE) for the Y_{est} representing the standard deviation of the group of scores is most useful in setting confidence limits that express a willingness to accept a regretted decision (Draper & Smith, 1998; Parker et al. 2010).

Summarizing Trend

Drawing upon methods from business, physical sciences, and biosciences, the Y_{est} , SEs and confidence intervals (CIs) can also be used to reliably evaluate growth (Draper & Smith, 1998; Payton, et al., 2003; Schenker & Gentleman, 2001). Parker et al., (2010) suggest that a method known as “approximate repeats” identified in Draper & Smith (1998) provides additional summaries of improvement necessary for high-stakes

decisions. Time in this method is modified by entering scores from adjacent weeks separately at one point in time (e.g., scores at week 2 and 4 are all entered at week 3) as shown in Figure 3.2a. Figure 3.2a shows a student's scores over 20 weeks of PM.

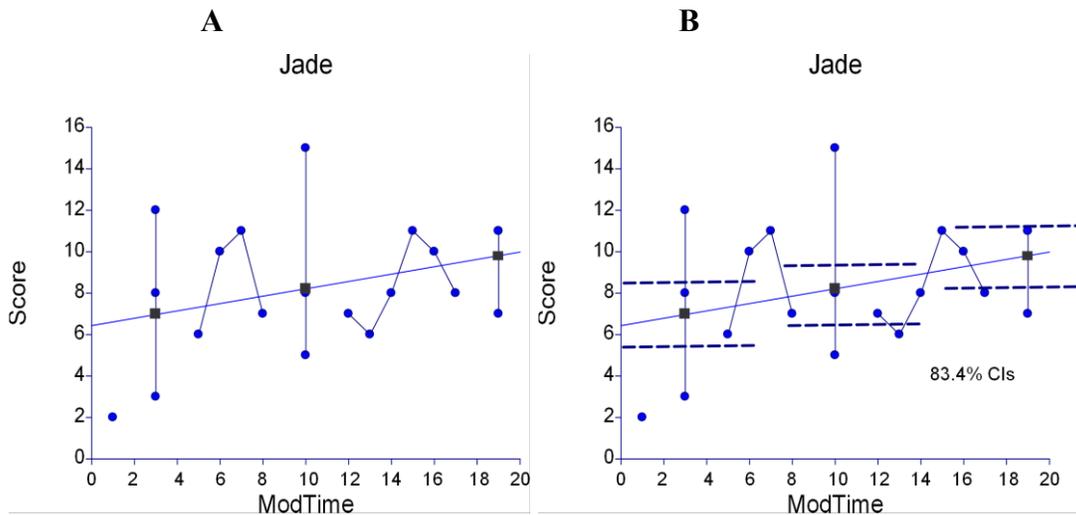


Figure 3.2. Static performance analysis of sample data set from 20 weeks of progress monitoring for repeated math concept and application probes adjacent to weeks 3, 10 and 19 **(a)** and confidence limits at 83.4% around Y_{est} scores **(b)**.

Scores at week 3, 10 and 19 were modified, so that scores from adjacent weeks were moved. In this example, scores from weeks 2 and 4, 9 and 11, and 18 and 20 were moved to week 3, 10 and 19 respectively. The Y_{est} scores indicated by the squares on the trend line provide static performance results at weeks 3, 10 and 19 instead of median scores or averages typically presented in the literature. A Y_{est} score, the best estimate of student performance at that time is part of the normal output in linear regression programs. The standard error (SE) for the Y_{est} accounts for more sources of error than

those accounted for in the CTT model (Christ & Coolong-Chaffin, 2007; Draper & Smith, 1966; Franklin, Allison & Gorman, 1996; Neter et al., 1996).

Reliable improvement between two points in time can be obtained, according to literature outside the social sciences, by using the SEs to calculate confidence intervals about the Y_{est} and examining for the presence of overlapping confidence intervals (Payton et al., 2003; Schenker & Gentleman, 2001). Methods from Payton et al. (2003) and Schenker and Gentleman (2001) are illustrated in Figure 3.2 B, which shows confidence intervals at 83.4% around the Y_{est} at weeks 3, 10, and 19. For important decisions, 90% to 95% CIs are standard (Nunnally 1967; Salvia & Ysseldyke, 2003) leaving a reasonable range of a 5 to 10 percent likelihood of error. Another way to limit error to 5% or less is to set CIs of 83.4% around two point estimates with similar SEs, as Payton et al. (2003) found in their investigation. The 83.4% CIs around the point estimates can be examined, according to Payton et al. (2003) for the presence of overlapping confidence intervals. According to Schenker and Gentleman (2001) this CI overlap method is equivalent to a Z or a t-test between two Y_{est} scores. Their study concluded that the confidence overlap procedure was equivalent to a statistical significance test at the $\alpha = 0.05$ level. In Figure 3.2B the presence of overlapping CIs set at 83.4% indicates with 95% certainty that real improvement did not occur between weeks 3 and 10 or weeks 3 and 19.

Another way to summarize growth reliably according to Draper & Smith (1998) and Neter et al. (1996) is to obtain a confidence band around the entire regression line as shown in Figure 3.3. In Figure 3.3 the confidence interval is set at 95% indicating less

than a 5% chance that the slope resides outside the confidence bands. Note how the confidence bands are narrow toward the center and wider on the ends. Draper and Smith (1998) explain that because the prediction is based on the full dataset one should expect that best prediction is closest to the middle of the range of scores and less towards the beginning and end of the range. Parker et al. (2010) also illustrated how confidence limits when calculated around the slope at a 95% level, as shown in Figure 3.3, include the Y_{est} and 83.4% CIs.

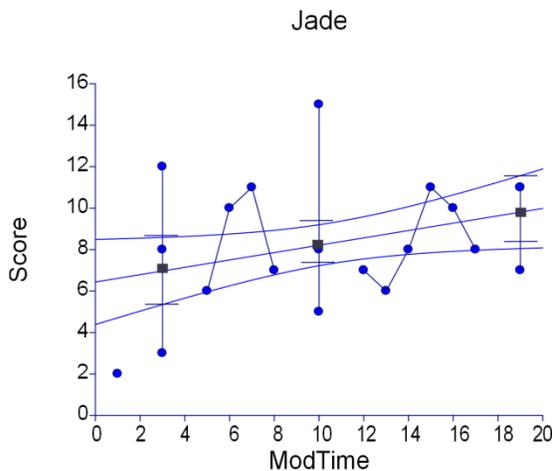


Figure 3.3. Confidence levels of 83.4% around the Y_{est} at week 3, 10 and 19, and 95% CIs about the slope.

Based on the progress monitoring literature one can conclude that PM measures have been vetted in the research in terms of being reliable and valid measures for instructional decision-making (low-stakes). In terms of reliably judging growth for more important, potentially irreversible or regrettable decisions, recent studies indicate that

CTT-based reliabilities do not sufficiently account for error. This is evident in several studies of oral reading fluency PM measures (Derr & Shapiro, 1989; Derr-Minneci & Shapiro, 1992; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008; Hintze & Christ, 2004; Hintze, Daly & Shapiro, 1998; Hintze, Owen, Shapiro & Daly, 2000; Jenkins, Graft & Migloretti, 2009; Poncy, Skinner & Axtell, 2005; Schatschneider et al., 2008). The literature has not identified issues with measurement error outside of oral reading fluency probes, and a paucity of articles provide methods to analyze and summarize PM results for important decisions (Deno, Fuchs, Marston, & Shin, 2001; Christ & Coolong-Chaffin, 2007; Fuchs, et al., 1993; Parker et al., 2010).

The present study is the first of a two-part study to examine the capacity of PM measures to reliably and precisely estimate growth using a math PM measure. The focus of this portion of the study is primarily on assessing growth from static performance levels to reliably estimate growth over time (Draper & Smith, 1998; Parker et al., 2010). The methods applied in this study are from fields outside of education to reliably estimate performance change necessary for making medium- to high-stakes decisions (Parker et al., 2010). These methods were applied to a large AIMSweb PM data set in mathematics. The study will answer the research question: Within an OLS regression model, what span of time is required to measure reliable change in performance (static measure of progress) from Y_{est} scores and their confidence intervals.

Method

An examination of the ordinary least squares (OLS) regression model with “approximate repeats” method was applied to an extant data set obtained from

AIMSweb. The data was limited to third grade progress monitoring results from the Mathematics Concept and Application (M-CAP) measure. Point estimates were calculated along with their standard errors and confidence intervals to determine how much time was needed to reliably assess growth.

Instruments

AIMSweb, a progress monitoring (PM) system, was the source for the data used in this study. It is a computer-based application used to measure progress in reading, mathematics, spelling, and writing providing practitioners with benchmark and progress monitoring measures. PM results were requested from M-CAP results for third grade. The system contains 30 M-CAP probes for progress monitoring at each grade level. Each eight-minute timed probe contains 29 problems. According to the online Administration and Technical Manual (NCS Pearson, 2009), the measures are designed to assess a broad array of mathematical concepts including number sense, operations, algebra, geometry, patterns and relationships, statistics and probability. These mathematical domains represent the core standards identified by the National Council of Teachers of Mathematics (NCTM) and proficiencies necessary to equip students for life-long mathematical thinking (NCTM, 1980; NCTM, 2006). Examiners are directed to administer the measure to the entire class and to encourage students to answer as many problems as possible in the time frame allotted. Test items are scored as correct or incorrect; partial credit is not given. Practitioners are guided to use their professional judgment in instances where the answer is correct but deviates from the answer provided

in the answer key. The primary goal is to ascertain if the student's response reflects an understanding of the task.

Internal consistency to examine the intercorrelation of test items was evaluated using Cronbach's alpha and split-half reliability. Probes had to meet a minimum CTT-based reliability standard of 0.80 or greater acceptable for benchmark assessments. Probes administered to 965 third graders had a reliability coefficient of 0.81, a group mean score of 20 items correct and a pooled standard deviation of 10.2. The SEM_{eas} was 3.64. Probes were sorted by the mean score and those scores that deviated the most from the mean were eliminated from the pool. A 95% confidence interval was selected along with the standard error of measurement to determine that the final selection of probes was statistically equivalent within each grade level.

Participants

The original data set, obtained through a written request by the researcher, included 535 individual third grade student score sets from the M-CAP. Students repeating third grade were excluded from the sample. The permission was reviewed and accepted by the PM system. Personally identifiable student information was removed and replaced with numerical codes and sent to the researcher in an Excel spreadsheet. The participants represented 91 school districts, 131 schools, and 163 classrooms from across the United States. Demographic information i.e., the location of the districts, gender, ethnicity was not provided to the researcher. From this database 92 participants were retained for this study using procedures described in the procedures.

Procedures

The study used a fixed time series design across 21 progress monitoring weeks for 92 third graders representing a national sample. The following procedures were implemented to select participants and analyze results with illustrations provided to describe methods based on Parker et al., 2010 that are rarely applied to school data.

- 1) A written request was sent to Pearson to obtain a large representative national data set of approximately 500 individual third graders' math concepts and applications progress monitoring results. Participant selection from the original dataset of 535 third graders used the following selection criteria. The participant had to have 21 weeks of progress monitoring data and have met the criteria identified by Shinn (1989) and Ardoin & Christ (2008) for being a "low performer" using baseline scores. Using the first three scores as "baseline" data, participants were included in the study if their median scores were less than 71% accurate; having fewer than five items correct and placing the participant's ranking at or below the 25th percentile based on nationally normed fall benchmarks for this measure (NCS Pearson, 2009).
- 2) Slope was calculated by entering weekly scores for each participant into a statistical package (e.g. NCSS) with a column for time and corresponding column for scores. A linear regression analysis was run on each student's score with output results that include slope and an R^2 index.
- 3) A measure of static score, the Y_{est} , was calculated to estimate student performance at weeks 3, 10 and 20. The Y_{est} was calculated within the statistical

package using a method known “approximate repeats” (Draper & Smith, 1998). One column labeled “Mod Time” contained modified times around weeks 3, 10 and 20. For example, Time from adjacent weeks around week 3 (e.g. week 2 and 4) was modified to the median value (to week 3). This procedure was repeated at week 10 and week 20. A second column containing the student’s scores was also created. Reports were requested in the system for the predicted Y means and predicted Y confidence level (estimated Y CL). The purpose of the estimated Y CL will be described in step 5.

- 4) Confidence intervals were calculated around the Y_{est} scores using the formula $b_1 \pm Z(1 - \alpha/df; n - df) s\{b_1\}$ to estimate a student’s growth over time (Neter, et al., 1996). CIs were also obtained from the statistical output in a linear regression program as described in step 3. An alpha level of .166 was preset for a confidence interval of 83.4%. An 83.4% CI is recommended by researchers outside the social sciences needed to reliably judge improvement between two point estimates with 95% certainty (Payton et al., 2003; Schencker & Gentleman, 2001).
- 5) Estimated Y confidence levels of 83.4% (set in the statistical package at an alpha level of 0.166) were calculated to examine growth between point estimates (Y_{est} s).
- 6) Differences between group means (the Y_{est} and SEM_{eas} at week 3 and the Y_{est} and SEM_{eas} at week 10 and the Y_{est} and SEM_{eas} at week 3 and the Y_{est} and SEM_{eas} at week 20) were compared using a two-tailed t-test. P-values were reported along

with confidence intervals to determine if there was significant change between weeks 3 and 10 and weeks 3 and 20.

Results

This study examined an alternate method known as “approximate repeats” to determine if this method could reliably assess growth needed for high-stakes decisions. Mean values and other descriptive data are presented in Table 3.1. Specifically, the study answered the research question, “Within an OLS regression model, what span of time is required to measure reliable change in performance (static measure of progress) from Y_{est} scores and their confidence intervals?” Results summaries for the study and referenced in this section are provided in tables in Appendix A.

In this study growth was assessed as per the literature by looking at slope using an OLS model. Slopes on the third grade AIMSweb Math Concept and Application Measure (M-CAP) were calculated using a linear regression program for the 92 third graders and reported in Table 3.1 along with trendedness indices and standard errors for the slopes. Slope values ranged from -0.01 to 1.26 with a standard deviation of 0.25 items per week. The average slope within this sample was 0.38 indicating the average student among these participants would be expected to improve his or her score on the M-CAP by approximately one point every two to three weeks. Trendedness indices or R^2 , a measure of score variation, are also reported in Table 3.1. The R^2 values for the participants in this study ranged from 0.00 to 0.91. The average R^2 index was 0.46 and the standard deviation was 0.23. An index of 0.46 indicates moderate variability and is interpreted as 46% of score variability can be explained by linear improvement.

In addition to looking at the slope and the stability of the slope, other data was examined in the OLS model that provides an indication of precision. In this study the SEslopes, reported in Table 3.1, ranged from 0.02 to 0.18 among the participants. For instance, the first student in Table 3.1 had a slope of 0.19 and a SEslope of 0.07. The standard error of the slope (SEslope) was reported at the 95% confidence level to more precisely estimate upper and lower limits within which the slope likely resides. These results indicate that the bands of confidence around the slope are at the upper limit of 0.33 and the lower limit of 0.05. One can interpret these results as, “We are 95% confident that the weekly rate of improvement for this student is between 0.05 and 0.33 items correct per week.” Confidence bands around static scores can also be reported using OLS with the “approximate repeats” method for reliably interpreting growth. In this study datum obtained from the OLS model were used to determine how many weeks are needed to reliably estimate growth.

OLS Analysis with Estimated Static Scores for Reliably Interpreting Growth

The method of “approximate repeats” provides performance results for static performance which can be used to estimate growth over time. Results from the OLS model with the “approximate repeats” method, reported in Table 3.2, were used to determine the span of time needed to reliably measure change in performance among the group of “low performing” participants. Results for this method were calculated in a regression package. The Y_{est} reported in Table 3.2 was calculated at week 3, 10 and 20 for each participant. The regression package also provided the standard errors of the Y_{est} and their associated CIs also reported in Table 3.2. The results were interpreted as

follows. Among the participants in this study, performance levels at week 3 ranged from 0.19 to 7.76 items correct. By week 10, the Y_{est} scores ranged from 2.16 to 17.39, and by week 20, the Y_{est} scores ranged from 2.66 to 29.66. Using these scores, ten students' scores changed little from week 3 to week 20 as in one example where the student's Y_{est} scores were 4.11 at week 3, 4.14 at week 10, and 4.18 at week 20. Other student scores changed dramatically as in another example where the student's scores doubled from 3.08 at week 3 to 6.76 at week 10, and 12.02 at week 20.

The static scores in the “approximate repeats” method were used to evaluate individual student growth over time by using the SE for the Y_{est} to establish confidence intervals (CIs) needed to reliably estimate growth from one point in time to another point in time. CIs were reported in Table 3.2 around each Y_{est} score to examine which students made reliable improvement between week 3 and 10 and weeks 3 and 20. An individual was judged to have made reliable improvement if confidence intervals did not overlap (Schenker & Gentleman, 2001). The results in Table 3.2 show that the 83.4% confidence intervals overlapped in 31 out of the 92 cases (or 34% of participants) between weeks 3 and 10 meaning that one can be 95% confident that real improvement occurred for 61 students as represented by the boldface type. The presence of overlapping confidence intervals was again examined between weeks 3 and 20. By week 20, confidence intervals did not overlap in 89% of the cases. In these instances we can be 95% confident that between weeks 3 and 20 growth for 82 students represents true improvement over that time period.

Since nearly one quarter of the students did not show reliable improvement between weeks 3 and 10, the presence of overlapping confidence intervals was extended to compare week 3 to subsequent weeks beginning with week 11 to determine if reliable improvement was made earlier than week 20. Table 3.3 provides the results of the 23 students who did not have reliable improvement at week 3 but did have reliable improvement at week 20. The results indicate that 15 more students demonstrated reliable improvement by week 13 and that by week 16 all but one participant in this group had made reliable improvement.

The visual technique of examining the presence of overlapping CIs, calculated around the Y_{est} scores at selected weeks using the “approximate repeats” method, attempts to replicate a t-test for comparing slopes between weeks 3 and 10 and 3 and 20. For comparative purposes a two-tailed t-test was run for each participant. Table 3.4 includes t-test results with 44 degrees of freedom. At a $p\text{-crit} = \leq .05$, the t-value must be 2.02 to reject the null hypothesis, that the means (the Y_{est}) are the same. Table 3.4 also provides t-values and p-values calculated for each mean, as well as, 95% CIs computed from the standard error of the difference between the Y_{est} scores. Results replicate the overlapping CIs results. Using the t-test 61 of the participants showed statistically significant growth by week 10 as indicated by t-ratios that exceeded the minimum ratio needed to be statistically significant at the $p = \leq .05$ level. When these t-test results were compared with the significance testing using overlapping confidence intervals, the results were the same.

Discussion

The purpose of this study was to examine data useful in judging the amount of time needed to reliably assess growth on the Mathematics Concept and Application (M-CAP) measure. This study, as suggested in the literature, applied the ordinary least squares (OLS) regression model but with the “approximate repeats” method to analyze progress monitoring (PM) results based on static scores (Christ & Coolong-Chaffin, 2007; Draper & Smith, 1998; Parker et al., 2010). In this study the OLS model with “approximate repeats” yielded point estimates and confidence intervals useful for reliably and precisely assessing growth. After 10 weeks of progress monitoring, two-thirds of the participants had made reliable gains, and after 16 weeks 95% of the participants had made reliable improvement.

If PM results are being recommended for important decisions (i.e., increased services or placement) the methods in this study provided static scores needed to reliably summarize results (Christ & Coolong-Chaffin, 2007; Parker et al., 2010). Overall, both performance level and trend are used for decision-making within the response to intervention (RtI) model (Fuchs & Fuchs, 1998). Contrary to the literature, which recommends the use of median scores or averages this study used a point estimate, the Y_{est} , to summarize static performance (Fuchs & Fuchs, 1998; Shinn, et al., 1989). The Y_{est} and other indices were obtained from the OLS model with “approximate repeats” indicating the stability, reliability and precision growth over time.

Throughout the PM literature the extent to which PM results were considered reliable has been determined through technical adequacy studies applying Classic Test

Theory (CTT). More recently, a small number of studies have shown that CTT-based reliabilities are insufficient for determining the reliability of PM results (Derr & Shapiro, 1989; Derr-Minneci & Shapiro, 1992; Dunn & Eckert, 2002; Francis, et al., 2008; Hintze & Christ, 2004; Hintze, et al., 2000; Jenkins, et al., 2009; Poncy et al., 2005). Rather than relying on median scores and averages typically reported in the PM literature (Foegen et al., 2007; Wayman et al., 2007) this study used indices to reliably report growth.

In this study, trendedness indices (R^2), also obtained from the OLS model, indicated the stability of a slope. Half of the participants had R^2 values that were moderate to high indicating that the improvement rates were moderate to very stable around the trend line. An R^2 value which explains the percentage of data that contributed to improvement supports decision-making in terms of determining the reliability of the trend line.

To precisely estimate how many weeks were needed to reliably estimate growth, this study applied the OLS model with “approximate repeats” method. Results of this study demonstrated that point estimates (Y_{est}) from a linear regression program could be used to reliably estimate growth between within 10 to 16 weeks for 95% of the participants. As stated in the literature the Y_{est} is a better estimate of performance at a point in time because it lies on the regression line, accounts for additional sources of error and is a mean score based on all of the data points in a time series (Christ & Coolong-Chaffin, 2007; Neter et al., 1996; Parker et al., 2010). A Y_{est} is a more reliable

estimate of performance than median scores because it resides on the slope (Christ & Coolong-Chaffin, 2007; Neter, et al., 1996).

This study also obtained indices needed to summarize results more precisely according to the literature (Christ & Coolon-Chaffin, 2007; Parker et al., 2010). To add precision to the summaries, confidence limits were set around point estimates at 83.4% as suggested by Schencker and Gentleman (2001) to examine the presence of overlapping confidence intervals. Confidence intervals did not overlap in two-thirds of the results between week 3 and week 10 indicating statistically significant change at the $p < 0.05$ level and identical to two-tailed t-test at the $p 0.05$ level. Since a full third of the cases did overlap, an additional procedure was added to examine at what point, prior to week 20, did statistically significant growth occur? The results indicated that 95% of the participants showed real growth between weeks 10 and 16 reflecting the sensitivity of the M-CAP to measure growth within a semester.

In sum the methods applied in this study strengthen the case for using other data in improvement summaries beyond reporting median scores and slopes when high-stakes decisions are being considered. However, it should be noted that there were limitations to this study to address in future research. The participants in this study represented a national sample, but the extant dataset was limited and presents threats to the internal validity of the study. In terms of selecting “low performers” for this study the researcher was limited to the information provided in the dataset, which included only AIMSweb results. Historical information about the participants, other than students in the sample had not been retained, was not obtained. Results from other state and local assessments

might have substantiated the identification of “low performing” in this study. In some instances notes were included in the dataset indicating if a student was receiving supplemental services or had been referred for special education. This information was not considered in the study because the field is not required in the AIMSweb system, it was completed inconsistently within the dataset, and there was no way of confirming the reliability of the information provided. Also the dataset did not include other information typically reported to assure readers of selection bias such as gender, ethnicity or socio-economic factors. This last issue presents limitations in terms of the generalizability of this study. The researcher had no way of confirming how the results of this study would differ given different subjects. This study also did not take into account treatment effects and how that might have influenced the results.

Future research should continue to examine ways to analyze and summarize PM results for important decisions. While the results of this study showed that reliable growth could be obtained in 10 to 16 weeks applying the “approximate repeats” method, this study should be replicated and ideally compared with other methods. The field would also benefit from identifying variables that influence growth. Further examination of the results in this study leave lingering questions about why some students made large gains over 21 weeks of PM and others showed almost no growth. When do the results suggest the presence of a disability? How can decision-makers be assured that treatments were delivered with fidelity and intensity? Are teachers evaluated based on the merits of their students’ PM results? These questions present larger issues that should undoubtedly fuel future debates and research in the area of PM for high-stakes decision-making.

CHAPTER IV

A COMPARISON OF MODELS FOR RELIABLY ESTIMATING GROWTH ON MATHEMATICS CONCEPTS AND APPLICATION MEASURES

Introduction

This chapter presents the second half of a two-part study on the capacity of progress monitoring (PM) measures to reliably assess growth. PM as described here refers to repeated measurement using curriculum-based assessments (CBM). Part one of the study examined techniques to reliably evaluate growth based on static performance levels. A technique known as “approximate repeats” (Draper & Smith, 1998) was applied in part one of the study, and results indicated that the Y_{est} was a more reliable estimate of student performance among the group of 92 “low performers” at a point in time. The Y_{est} was also useful in summarizing growth over time for each participant. In part one of the study, two-thirds of the participants through visual inspection of overlapping confidence intervals reliably improved in ten weeks and 95% improved within 16 weeks of PM. In this paper, the second half of the study examined parametric and nonparametric analytical models and compared slopes and other indices needed to reliably report growth among “low performing” students. This part of the study also focused on Stage 2 PM research, which has been less represented in the literature according to Fuchs (2004) and two recent meta-analyses of PM studies (Foegen, Jiban, & Deno, 2007; Wayman, Wallace, Wiley, Ticha, & Espin, 2007).

Growth from progress monitoring (PM) in the PM literature is typically assessed in terms of slope (Deno 1985; Deno, 2003; Deno, Fuchs, Marston & Shin, 2001; Foegen et al., 2007; Fuchs, Fuchs, Hamlett, Walz & Germann, 1993; Wayman, et al., 2007). Practitioners should be accustomed to looking at graphed PM results and determining if trend lines are moving in a positive direction, adjusting instruction if growth is not apparent. Practitioners should be familiar with summarizing slope as in this example, “Tyler’s weekly rate of improvement on third grade math computation skills is three digits per week since he was assigned to a peer tutoring group for 30 minutes daily. At this rate we expect he will be performing at the 50th percentile by the end of the semester.” For important decisions such as using the PM results for determining program changes, other analytic techniques are being recommended but are less well known (Christ & Coolong-Chaffin, 2007; Parker, Vannest, Davis, Clemens, 2010). These techniques are meant to address some technical features of slope so that improvement summaries are more reliable and precise (Christ & Coolong-Chaffin, 2007; Hintze & Christ, 2004; Parker, et al., 2010; Poncy, Skinner, & Axtell, 2005).

Technical Features of Curriculum-based Measures (CBM) for Progress Monitoring

Technical features of PM measures to assess growth are primarily reported in reliability coefficients and slope in two stages of research (Fuchs, 2004). The technical features were recently summarized in two meta-analyses; one in reading and one in mathematics with slope only addressed in approximately one-third of the studies (Fuchs, 2004; Foegen et al., 2007; Wayman et al., 2007). Growth or improvement rates from PM results, according to some studies of oral reading fluency measures, are unstable due to

several factors: 1) differences in testing environments (Derr & Shapiro, 1989; Derr-Minneci & Shapiro, 1992; Hintze & Christ, 2004; Hintze, Daly & Shapiro, 1998; Jenkins, Graff & Miglioretti, 2009); 2) developmental differences (Hintze et al., 1998; Poncy, et al., 2005); and 3) measurement difficulty (Dunn & Eckert, 2002; Hintze, et al., 1998; Hintze, Owen, Shapiro & Daly, 2000; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008). Hintze & Christ (2004) found that uncontrolled measurement error is compounded across multiple administrations and inflates performance results. These issues affect the reliability of PM results (Wayman, et al., 2007).

Research in mathematics PM measures and their capacity to measure growth is significantly smaller than the research of reading PM measures and smaller still if one is interested in measures that assess conceptual understanding and application to real world problem solving (Foegen et al., 2007). Of the 32 studies of mathematics progress monitoring measures, Foegen, et al. (2007) identified only four studies that evaluated growth on mathematics concepts and applications measures (Fuchs, Hamlett, & Fuchs, 1999; Fuchs, Fuchs, Hamlett, Thompson, Roberts, Kupek, & Stecker, 1994; Helwig & Tindel, 2002; Shapiro, Edwards, & Zigmund, 2005). Three of the four studies examined growth using the Monitoring Basic Skills Progress (MBSP) (Fuchs, et al., 1999; Fuchs et al., 1994; Shapiro et al., 2005) and in one study the researchers developed a set of probes to progress monitor students in eighth grade (Helwig & Tindel, 2002). The reliability of the measures to assess growth, as reported in three of the four studies, was evaluated using CTT-based alternate form reliability. Helwig and Tindel (2002) reported reliability coefficients from 0.81 to 0.88 on four probes used in their study. Fuchs et al. (1994)

correlated MBSP probes by grade level obtaining coefficients in grades 2, 3, and 4 at 0.98, 0.94, and 0.97 respectively. Fuchs et al. (1999) tested the reliability of the MBSP measures for grades 5 and 6 obtaining coefficients of 0.97 at both grade levels.

The technical features of slope, to examine growth over time, were reported in all four studies of student performance on concept and application measures (Fuchs, et al., 1999; Fuchs, et al., 1994; Helwig & Tindel, 2002; Shapiro, et al., 2005). Three studies reported slopes for the MBSP Concepts and Applications measure (Fuchs, et al., 1994; Fuchs, et al., 1999; Shapiro, et al., 2005). Fuchs, et al. (1994) calculated slopes for 140 students in second through fourth grade in an urban school district in the southeast. The participants were primarily general education students with the exception of 12 students with learning disabilities. Average slopes obtained in this study were 0.40, 0.58, and 0.69 in grades 2, 3, and 4 respectively. Fuchs, et al. (1999) computed slope averages for 51 fifth graders and 44 sixth graders on the MBSP. Slopes at grade 5 were 0.38 and at grade 6 were 0.26 items correct per week. The MBSP was used in a study conducted by Shapiro et al. in 2005. This study examined growth among 109 students with high incidence disabilities receiving special education services in grades 2 through 5. The MBSP was administered across the school year at the student's performance level based on teacher evaluation. This means a student may have been in fifth grade but was administered PM measures at third grade level. This study contained a training component for teachers to use data-based decision rules to adjust instruction based on student results. The participants' range of performance was between -0.5 to 1.50 items correct per week in this study on the MBSP concept and application measure. The

average slope for all of the participants was 0.38 items correct per week, and by grade level average slopes were 0.36 at second grade, 0.37 at third grade, 0.44 at fourth grade, and 0.52 at fifth grade.

Despite decades of research on the technical features of PM measures, there are significantly fewer studies in mathematics and calls for future studies on the technical features of slope and the reliability of the measures to inform high-stakes decisions (Christ & Coolong-Chaffin, 2007; Foegen et al., 2007; Parker et al., 2010; Wayman et al., 2007). These scholars call for new standards and skills to support the field shifting PM use for low-stakes decisions (informing instructional changes for students on IEPs) to high-stakes decisions (informing services or placement changes). They question if PM slopes are reliable for high-stakes decisions. Foegen et al. (2007) recommend reliabilities of 0.90 or higher for high-stakes decisions. Christ and Coolong-Chaffin (2007) and Parker et al. (2010) indicate that CTT-based reliabilities are inappropriate for PM results. For high-stakes decisions these authors assert that PM results be analyzed using ordinary least squares (OLS) regression models and include error and confidence limits. The OLS model and related indices are recommended for providing more reliable and precise analyses of PM results (Christ & Coolong-Chaffin, 2007; Good & Shinn, 1990; Parker, et al., 2010; Shinn, Good, & Stein, 1989). Alternate models, suggested for students whose profiles exhibit less growth or nonlinear growth over time, have also been explored in single-case research (Parker, Vannest, Davis, & Sauber, 2011). A brief summary of models, the parametric OLS model and nonparametric models Theil-Sen

and Tau-- new to school data, will be useful in terms of comparing these models and providing practitioners with a nonparametric alternative to the OLS model.

The Ordinary Least Squares Model

OLS is one of several analytic techniques presented in the literature useful for assessing growth in time series data for individuals (Gorman & Allison, 1996; Manolov, Solanas, Sierra, & Evans, 2011; Parker, Vannest & Davis, 2011; Parker, et al., 2010; Parker, Vannest, Davis & Sauber, 2011). The OLS model has routinely informed data-based decisions in such fields as business and health (Draper & Smith, 1967; 1998; Neter Kutner, Nachtsheim, & Wasserman, 1996). The regression model for PM describes the statistical relationship between PM scores and the variable of time which is constant and serves as a predictor variable (Neter et al., 1996). The strength of the relationship between time and score is determined by estimating the squared distance of scores from the regression line. The PM literature applies OLS to analyze growth, but summaries are limited only to slope (Foegen et al., 2007; Fuchs & Fuchs, 1998). For example the student in figure 4.1 has a slope of 0.43 indicating an improvement rate of approximately one point correct every two to three weeks.

There are a number of other indices that are part of the routine output in linear regression programs that inform practitioners of the variability of time series data to adequately summarize results for important decisions (Christ & Coolong-Chaffin, 2007; Parker et al., 2010). For instance, trendedness indices such as an R^2 index provided in linear regression programs indicate the percent of scores that follow the trend line. Scores with a lot of variance or distance from the trend line result in low R^2 indices.

The graphed scores in figure 4.1 do not closely follow the trend line resulting in an R^2 index of 0.23, which is a weak trend index indicating that only 23% of the scores

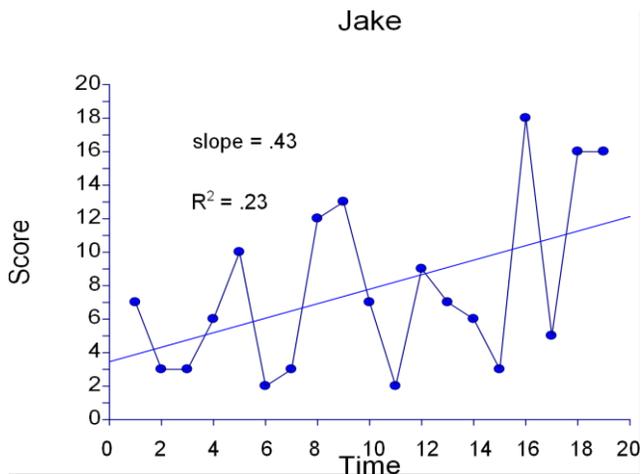


Figure 4.1. A graph of one participant's scores across 20 weeks of PM that includes slope and R^2 index as calculated in a linear regression program.

contributed to linear improvement. The influence of error on the slope is captured in the standard error of the slope (SEslope). The SESlope is used to calculate confidence intervals (CIs) within which the slope resides. Scholars within education are recommending practices from medicine in situations where the decision-making is irreversible (Christ & Coolong-Chaffin, 2007; Djulbegovic, Schwartz, & McMasters, 1999; Parker et al., 2010). For important decisions the SESlope and CIs provide more precise estimates of performance needed for medium- to high-stakes decisions (Christ & Coolong-Chaffin, 2007; Parker et al., 2010). For instance an evaluation team would want to have a high level of confidence (e.g., confidence

level set at 95% indicating their willingness to accept an error in 5% or fewer cases) if examining PM growth rates for the purposes of determining eligibility.

OLS is still considered to be a superior model to other techniques, because growth in PM results is linear and primarily monotonic (meaning that as time increases the mean value of Y, scores, increases) (Deno, 2003; Fuchs, et al., 1993; Sprent, 1993). A number of researchers have criticized the use of OLS with time series data because the data can violate parametric assumptions of constant variance, normality, and linearity (Cohen & Cohen, 1983; Gorman & Allison, 1996; Neter et al., 1996; Parker & Brossart, 2003). Tests of assumptions, which are normal outputs in linear regression programs, support practitioners in identifying if the results adhere to linear assumptions. These tests include the Shapiro-Wilk test of normality used to assess how well the slope fits a regression line. High correlation coefficients are indicative of normality (Neter et al., 1996). The modified Levene test is used to determine whether the error terms have constant variance (Neter et al., 1996). This is indicated by small error terms. Another test, the Durbin-Watson test of autocorrelation, examines if the error terms are autocorrelated, indicating interdependence of datapoints (e.g. predictability of data based on adjacent scores that is unrelated to trend) (Neter et al., 1996). If these tests show abnormal trend, a lack of constant variance or the presence of autocorrelation then the OLS model may not be appropriate for modeling growth for low performing students.

Nonparametric Analyses of Growth

Recognizing that change among low performers may not always be linear, growth can also be analyzed using nonparametric tests. Nonparametric analyses using Theil-Sen and Tau are new to education studies but are routinely applied in the physical

sciences, business, and health to model growth (Conover, 1980; Hollander & Wolfe, 1999; Parker, Vannest, Davis, and Sauber, 2011; Sprent, 1993). Theil-Sen slope is equivalent to LR slope and interpreted as rate of improvement (Hollander & Wolfe, 1999). Theil-Sen is a non-parametric analysis useful for analyzing datasets where results are not uniformly positioned around the trend line due to bounce or outliers (Weiss, 2003). Theil-Sen, analogous to linear regression slope, is a calculation of the median slope of data pairs within the set. Theil-Sen slope is calculated first by finding all possible slopes where N is all data points and calculated $[N*(N-1)]/2$ to provide all possible slopes. Next, slopes for all data pairs are calculated using the algorithm $\text{slope} = \frac{Y_b - Y_a}{X_b - X_a}$, the scores for data points “a” and “b” for all X and Y data pairs. The median value of all the mini slopes, the Theil-Sen slope, is given by $a_i = y_i - b*x_i$ (Sprent, 1993). Unlike linear regression in which the line of best fit is an estimated mean of the data pairs over time, Theil-Sen is a line that fits through the median of pairs of data.

In the OLS model, time is the constant and score is the dependent variable. In the Tau model, the nonparametric trend analysis is used to determine if time and score are independent (Hollander & Wolfe, 1999). Tau is a trendedness index similar to the R^2 in the OLS model. Tau and R^2 cannot be directly compared because Tau is interpreted as the percent of data, linear or otherwise, that show improvement. Therefore, this index is useful when data do not adhere to linear assumptions. It is also immune to outliers or data with a lot of variability and bounce (Hollander & Wolfe, 1999) and with respect to “low performers” may be a more robust indicator of growth. Kendall’s Tau is calculated by subtracting the number of pairs not showing improvement from the pairs that do

show improvement divided by the total number of pairs [# of positive pairs - # of negative pairs/ total pairs]. Distribution results from Parker, Vannest & Davis (2011) are useful for interpreting effect sizes or indicating that change occurred.

The literature indicates that the OLS model and the non-parametric Theil-Sen and Kendall's Tau provide pertinent data to inform important decisions. PM studies have not compared these models or examined their capacity to model growth for important decisions. This study is a comparison of two analytic techniques, Theil-Sen and OLS, from which results will be compared to examine the capacity of the models to yield reliable and precise summaries of improvement for "low performing" students. The study addressed the following research questions:

- 1) Among low performing students on the M-CAP, how closely do slopes when calculated with the Theil-Sen method approximate OLS regression slopes?
- 2) How reliable are the slopes as indicated by trendedness indices, standard errors and ratios of score/SEslope among low performing students on the M-CAP?
- 3) To what extent do data series among low performers fail to meet parametric assumptions of normality and equal variance?

Method

The methods used in this study were to compare the parametric ordinary least squares (OLS) model to nonparametric models using extant data made available through AIMSweb. The data set was limited to third grade progress monitoring results from the Mathematics Concept and Application (M-CAP) measure. Statistical analyses were

conducted to determine the reliability of slopes and other indices necessary to support high-stakes decisions.

Participants

The original data set, obtained through a written request by the researcher, included 535 individual third grade student score sets from the M-CAP. Students repeating third grade were excluded from the sample. The permission request for PM data was reviewed and accepted by the PM system. Personally identifiable student information was removed and replaced with numerical codes and sent to the researcher in an Excel spreadsheet. The participants represented 91 school districts, 131 schools, and 163 classrooms from across the United States. Demographic information i.e., the location of the districts, gender, ethnicity were not provided to the researcher. From this database 92 participants were retained for this study using procedures described later in the methods section.

Instruments

This study used the M-CAP progress monitoring measure from the AIMSweb system. The system contains thirty equivalent M-CAP probes for progress monitoring. Each eight-minute timed probe contains twenty-nine problems. According to the online Administration and Technical Manual (NCS Pearson, 2009), the measures are designed to assess a broad array of mathematical concepts including number sense, operations, algebra, geometry, patterns and relationships, statistics and probability. These mathematical domains represent the core standards identified by the National Council of

Teachers of Mathematics (NCTM) and proficiencies necessary to equip students for life-long mathematical thinking (NCTM, 1980; NCTM, 2006).

A protocol for administering and scoring each progress monitoring measure is described in detail in the online manual. Third graders are administered the measure class-wide and are encouraged to answer as many problems as possible in the eight-minute time frame allotted. Test items are scored as correct or incorrect; partial credit is not given. Practitioners are guided to use their professional judgment in instances where the answer is correct but deviates from the answer provided in the answer key. The primary goal is to ascertain if the student's response reflects an understanding of the task.

Probes were developed using the widely recognized and accepted standards from the NCTM (NCS Pearson, 2009; NCTM, 2006). Several procedures were used to develop equivalent measures and assess for reliability and validity. First, test items were developed by professional test developers and experts in the field of mathematics and field tested. Anchor probes were then made for each grade level and used to create equivalent measures in terms of proportion of item type, difficulty and item placement. Reviewers comprised of mathematics teachers and content experts reviewed the probes using a rubric for such qualities as the appropriateness of the items for the grade level assessed, vocabulary, mathematical symbols, question format, etc. According to the manual, the feedback was positive and suggestions for specific items were incorporated into the probe revisions. Three pilot studies were then conducted to determine if the

probes covered the desired content and upheld critical psychometric properties e.g. item completion time, alternate form reliability.

Internal consistency to examine the intercorrelation of test items was evaluated using Cronbach's alpha and split-half reliability. Probes met a minimum reliability standard of 0.80 or greater acceptable for benchmark assessments. Probes administered to 965 third graders had a reliability coefficient of 0.81, a group mean score of 20 items correct and a pooled standard deviation of 10.2. The SEM_{eas} was 3.64. Probes were sorted by the mean score, and those that deviated the most from the mean score were eliminated from the pool. A 95% confidence interval was selected along with the standard error of measurement to determine that the final selection of probes was statistically equivalent within each grade level.

Procedures

This study used a fixed time series design across 21 weeks of progress monitoring for 92 third graders representing a national sample. The following procedures were implemented to analyze results with illustrations provided to describe methods typically applied outside of education (Parker et al., 2010).

- 1) Participant selection from the original dataset of 535 third graders used the following selection criteria. The participant had to have 21 weeks of progress monitoring data and have met the criteria identified by Shinn (1989) and Ardoin & Christ (2009) for being a "low performer" using baseline scores. Using the first three scores as "baseline" data, participants were included in the study if their median scores were less than 71% accurate; having fewer than five items

correct and placing the participant's ranking at or below the 25th percentile based on nationally normed fall benchmarks for this measure (NCS Pearson, 2009).

The request was for a large representative national data set of approximately 500 individual third graders' math concept and application progress monitoring results.

- 2) Weekly scores for the sample were entered into a statistical package, e.g., NCSS, Stat-Plus, Stata, SAS, etc., with a column for time and corresponding column for scores. A linear regression analysis was run on each student's score with output results that include slope, an R^2 value, the standard error of the slope (SEslope), a correlation coefficient for testing normality, a test statistic for the modified Levene test of equal variance, and a test statistic for the Durbin-Watson test for autocorrelation.
- 3) Student scores were entered individually into a nonparametric statistical package (e.g., StatsDirect, Wessa, WinPepi) to calculate Theil-Sen slope.
- 4) Tau-U, a trendedness index to be compared with R^2 , and Z scores were calculated by entering individual student scores into a web-based calculator (e.g., <http://www.singlecaseresearch.org/calculators/tau-u>).

Results

In this portion of the study the parametric ordinary least squares (OLS) and nonparametric Theil-Sen and Tau-U were compared to address the remaining research questions:

- 1) Among low performing students on the M-CAP, how closely do slopes when calculated with the Theil-Sen method approximate OLS regression slopes?
- 2) How reliable are the slopes as indicated by trendedness indices, standard errors and ratios of slope/SEslope among low performing students on the M-CAP?
- 3) To what extent do data series among low performers fail to meet parametric assumptions of normality and equal variance?

Results summaries for the study, and referenced in this section, are provided in tables in Appendix B.

Research Question 1: Among low performing students on the M-CAP, how closely do slopes when calculated with the Theil-Sen method approximate OLS regression slopes?

One of the purposes of this study was to compare the utility nonparametric analyses to the parametric OLS model. Slope sizes were first compared between linear regression slope in the OLS model to its nonparametric counterpart the Theil-Sen estimator. Slopes in both models are analogous – indicating rate of improvement. Slope results for each test are reported in Tables 4.1 and 4.2. Slope size signals to practitioners the extent to which growth is occurring. Table 4.1 includes the slope ranges, means, medians and standard deviations from each model for the group. The descriptive results in Table 4.1 were used to broadly compare slopes calculated by OLS and Theil-Sen. Descriptive results in Table 4.1 show minor differences between the two models in terms of the range of slopes, slope means, median slope of each model and standard deviations of the slope in each model. Group results indicate slopes between the two models are

nearly identical. To more clearly articulate differences between the two models, individual results are listed in Table 4.2 including slopes and their standard errors.

Differences in slope size among participants are shown in Table 4.2 using shaded cells to indicate small (white), moderate (light gray) and large (dark gray) improvement rates over time. The results of the linear regression analysis show that rates of improvement (or slope) on the M-CAP ranged from -0.10 to 1.26 items correct per week. The median rate of improvement among the participants, as shown in Table 4.1, was 0.33 items correct. This means among this group, the typical low performing third grader requires approximately three weeks to improve by one data point. In this sample 7% of the participants had large improvement rates indicating that they gained approximately one point every week. Eighteen percent of the participants demonstrated moderate rates of improvement which translates to a one point gain every two weeks.

The Theil-Sen slopes in this sample range from 0.00 to 1.23 items correct per week and, all but six cases were nearly identical with expected weekly gains as calculated through OLS. Slope differences between the two models varied by 12 to 53 points in six participants' results in the moderate to large slope range. These six cases are noted in Table 4.2 with the slopes underlined for reference. In all cases the OLS model had the larger slope. No other distinguishing characteristics were observed that set these cases apart from other data.

Research Question 2: How reliable are the slopes as indicated by trendedness indices, standard errors, and ratios of slope/SEslope among low performing students on the M-CAP?

The reliability of the slopes in this study were evaluated in three ways, the 1) stability of slopes as measured by trendedness indices R^2 and Tau; 2) precision of slopes as indicated by the standard error of the slope and confidence intervals; and 3) power of the model to accurately model growth by comparing the ratio of slope/SEslope useful in evaluating statistical significance (Thompson, 2006). The results are reported in Tables 4.2 and 4.3 and described individually in the remainder of this section.

Stability

Reliable results should include some indication of the stability of the scores. In this study stability was examined by the trendedness indices R^2 and Tau. The trendedness index of R^2 is interpreted as the percent of data contributing to improvement, and Tau is interpreted as the percent of data that show improvement (Parker et al., 2011). Due to the differences in these indices they cannot be directly compared, but they do inform practitioners of the stability of the slope. In an OLS analysis, the R^2 index is an indication of the consistency with which scores contribute to growth over time. The R^2 indices as reported in Table 4.2 in the sample ranged from 0.00 to 0.91. An R^2 index of .91 is translated as 91% of the score variations can be explained by linear improvement. Large indices of 0.60 and greater were present in 35% of the participants' results. Another third of the participants had moderate indices, and a third of the cases had small indices of 0.39 or lower. Kendall's Tau is a non-parametric trendedness index, analogous to R^2 , and provides a rank correlation coefficient equal to

1. If in this study, all pairs of data for an individual show improvement, Tau is interpreted as the percent of data that show improvement (Conover, 1980; Parker et al., 2010; Sprent, 1993). Using this analysis, pairs of data that contributed to individual student improvement among the participants ranged from 0.03 to 0.84 with the latter having 84% of scores contributing to improvement over time. Using distribution results from Parker, Vannest & Davis (2011), half of these trend indices represent low effect sizes. The remainder of the participants' results were evenly distributed between either moderate or large effect sizes.

Precision

Table 4.2 also reports the SEslope and 95% confidence intervals for each participant to determine if those results provide more precise growth. The SEslope provides bands of confidence around the r coefficient providing a known likelihood that the true slope lies within the bands of confidence. The SEslope within the OLS model for individuals ranged from 0.02 to 0.18 and the Theil-Sen SEslope for individuals ranged from 0.00 to 0.29. The SEs in the Theil-Sen model were slightly larger for each participant than those calculated with OLS. Small error produces narrow bands of confidence. For example, in one participant's results the SEslope was 0.02. As a result we can say with 95% confidence that the student's score contributing to improvement was between 9% and 16%. More than three-quarters of the participants had wide bands of confidence and those did not differ widely between the parametric and nonparametric analyses.

The likelihood that the scores consistently follow a trend line in the OLS model or consistently improve over time in the Tau model is captured in the trendedness indices of R^2 or Tau. Another statistical output that informs the consistency is the p-value. P-values were reported in Table 4.6 to determine the likelihood that improvement was beyond chance levels. Based on the number of scores reported for each participant, results are statistically significant at $p < .05$. P-values are provided as normal output from statistical packages and web-based calculators. Consistent improvement over time was observed in more than 90% of the participants using the OLS model. In the Tau model, consistent improvement, not limited to linear improvement, was also observed in approximately 98% of the participants.

Power

Given that the OLS provides linear analysis and the Tau model analyzes growth linear or otherwise, which model demonstrates more power to sufficiently report growth? A ratio of slope to SEslope, equivalent to a Z or t-score, was computed to compare slopes from parametric and non-parametric analyses. The ratio of slope to SEslope was hand calculated and reported in Table 4.3, and the model with the largest ratio was considered to have more power. Both the median ratios for the OLS and Tau models, which were respectively 4.14 and 3.30, and ratios for each individual were compared. Medians for the model and three out of four individual ratios indicate the OLS has more power than the non-parametric model for summarizing growth.

Research Question 3: To what extent do data series among low performers fail to meet parametric assumptions of normality and equal variance?

Given that the participants in this study were selected because they were deemed “low performing,” this study sought to determine if the data failed to meet parametric assumptions of normality and equal variance, the focus of the third research question. The results to address this question are provided in Table 4.4 which includes the results of several tests used to examine parametric assumptions. The results were obtained as part of the normal output in the linear regression program. The Shapiro-Wilk test of normality was used to assess how well the slope fits the regression line. The correlation coefficients from the Shapiro-Wilk test of normality ranged from 0.66 to 0.99. W statistics of 0.96 and higher for $n=21$ indicate no evidence of non-normality (Shapiro & Wilk, 1965). Forty-two percent of the 92 participants had W statistics of 0.96 or higher indicating that the participants’ error terms were normally distributed (Neter et al., 1996). While the error terms were normally distributed, the modified Levene test was used to examine if the variance of the error terms was constant. At an alpha level of .05, the t test (.975; 19) would require $|tL| \leq 2.093$ to conclude that the error variance is constant (Neter et al., 1996). The test statistics ranged from 0.00 to 9.59 with 75% of the sample having error terms that were constant. Notably, 70% of the cases violated assumptions of either normality or equal variance. The final test of residuals was the Durbin-Watson test to determine if the error terms over time were autocorrelated. At a level of significance of .05, if the Durbin-Watson test was less than 1.42 for $n=21$, it would be concluded that error terms were positively autocorrelated (Neter et al., 1996). In this study, error terms were positively autocorrelated in approximately 20% of the cases.

Discussion

This study compared results obtained using ordinary least squares (OLS) regression to non-parametric analyses, to address issues of reliability and precision identified in the PM literature. Overall, differences were not found between the models and both models provided indices necessary for reporting the stability, reliability and precision of growth over time.

Specifically to answer the first research question, slopes, or individual improvement rates, between the two models were similar, varying only in how growth is interpreted within the models. The parametric OLS model describes the statistical relationship between PM scores and the variable of time, which is constant and serves as a predictor variable (Neter et al., 1996). The nonparametric counterparts, Theil-Sen and Kendall's Tau, do not require linear improvement and may be advantageous when PM results have slow growth or growth not readily apparent when examining slopes alone. One quarter of the participants had slopes with improvement rates that were moderate to large, indicating growth rates of one item correct every one to two weeks.

The second research question in this study sought to determine if the slopes were reliable using trendedness indices, standard errors and ratios of score/SEslope. Trendedness indices are one of several data sources needed to address measurement error issues associated with PM results. The trendedness index in the OLS model is the R^2 value which indicates the percent of score variation explained by time. The Tau index indicates the percent of improved scores without assumptions of linearity. Trendedness indices, Tau and R^2 cannot be directly compared, but they were a useful indicator of the

stability of the growth among the participants. In this study two-thirds of the participants analyzed using the OLS model had stable rates of improvement. One half using the nonparametric analysis had effect sizes indicating that a moderate to large number of data pairs were improved scores. Summaries of growth without an indication of the consistency with which the scores follow the trend line or improve can unduly influence conclusions about the results and provide no indication of the significance, precision or power of the model to accurately reflect growth (Kazdin, 1988; Parker et al., 2010).

A lack of consistency in the results indicates a lot of error, a current issue in the PM literature, so this study included reports of error and confidence intervals to estimate the reliability of the slopes. The researcher, consistent with other high-stakes assessments, established the confidence coefficient at 0.95 indicating that within multiple administrations the researcher was willing to accept error in 1 out of 20 administrations. In both models the standard errors produced wide confidence intervals meaning the true slope resided within a wide range. More forgiving confidence limits of 80% to 85% (typically applied to lower stakes decisions) would narrow the confidence band providing practitioners with a more precise estimate of growth. To support decision-making, p-values were also calculated to indicate if growth was beyond chance levels. Consistent improvement was observed in 90% of the OLS results and 98% of the nonparametric results indicating that p-values along with confidence intervals would be more useful in supporting decision-making.

The final analysis, in examining the reliability of slopes, was to determine which model, the parametric OLS model or the non-parametric model, more appropriately

growth for individuals. A ratio of slope/SEslope was used to directly compared models. This study found that the OLS model had more power in 75% of the cases to reliably estimate growth among these participants. However, 70% of all cases failed at least one of two tests of parametric assumptions. In light of the results of this study, the nonparametric model was not as powerful but yielded similar slopes, offered an indication of the stability of growth through the trendedness index, and did not have to adhere to parametric assumptions of normality. In sum, the nonparametric model was a more robust indicator of growth among this group of “low performing” third graders. Further research on this key finding is warranted given the results of this study.

There are several limitations that should be noted as they threaten the internal and external validity of the results in this study. The participants in this study represented a national sample, but the extant data did not include historical information about the participants other than they were not repeating third grade. The dataset did not include information typically reported to assure readers of selection bias such as gender, ethnicity or socio-economic factors. In terms of selecting “low performers” for this study the researcher was limited to the information provided in the dataset, which only included AIMSweb results. Results from other state and local assessments might have substantiated the identification of “low performing” in this study. In some instances notes were included in the dataset indicating if a student was receiving supplemental services or had been referred for special education. This information was not considered in the study because the field is not required in the AIMSweb system, it was completed inconsistently within the dataset, and there was no way of confirming the reliability of

the information provided. Limited participant information affects the generalizability of this study. There is no way of confirming how the results of this study would differ given different subjects. This study also did not take into account treatment effects and how that might have influenced the results.

The PM literature has only just begun to address issues with using PM results to inform medium- to high-stakes decisions. The available studies have been limited to issues with oral reading fluency PM measures and results. The study in this dissertation is the first, known by the researcher, that addresses methods for reliably analyzing mathematics PM results necessary for medium- to high-stakes decisions. It is also the first study known to the researcher that applies nonparametric Theil-Sen and Tau to PM results. More research is needed on the methods applied here to determine if results can be generalized to other participants and PM domains. The complexity of the analyses described in this study must also be considered. How are these methods presented to practitioners in a way that is more digestible to those who are not statisticians or graduates with advanced degrees in educational testing? What components of the procedures can be incorporated into PM systems in ways that becomes accessible and useable by practitioners? Overall, the methods applied in this dissertation must be vetted in the research. It is essential that there is more discourse on the limitations of the PM research to adequately address measurement error and the impact it has on the reliability and precision of PM results.

CHAPTER V

CONCLUSIONS

The purpose of this dissertation was to use methods within parametric and non-parametric analyses to more reliably and precisely summarize growth on the Mathematics Concepts and Applications (M-CAP) progress monitoring (PM) measure. The methods explored in this dissertation are rarely applied to school data but add to the emerging literature within the field examining the reliability of PM results.

The dissertation first presented a literature review describing the 1) development of curriculum-based measures for the purposes of progress monitoring; 2) research over two decades supporting the technical adequacy of the measures; and 3) policy shifts impacting PM use and data-based decision-making. A review of the literature shows the progression of PM from its primary function in supporting low-stakes decisions (e.g., in-class instructional decisions) to being recommended, more recently in an era of more accountability, for medium- to high-stakes including eligibility for special education (Deno, 1985; Foegen, Jiban, & Deno, 2007; Fuchs, & Fuchs, 1998; McGlinchey & Hixon, 2004; Restori, Gresham, & Cook, 2008; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Support for this shift comes from over two decades of PM research verifying the technical adequacy of the measures to assess growth at a point in time (Stage 1) and over time using slope (Stage 2), and the reliability of the results based on Classic Test Theory (CTT). Primarily within the last decade, a small number of studies show that CTT-based reliabilities are lower and less precise with PM results than

typically reported (Derr & Shapiro, 1989; Derr-Minneci & Shapiro, 1992; Dunn & Eckert, 2002; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008; Hintze & Christ, 2004; Hintze, Daly, & Shapiro, 1998; Hintze, Owen, Shapiro, & Daly, 2000; Jenkins, Graft, & Miglioretti, 2009; Poncy, Skinner, & Axtell, 2005). Others argue that reliable and precise estimates of growth are necessary if PM results are used for important educational decisions (Christ & Coolong-Chaffin, 2007; Parker, Vannest, Davis, and Clemens, 2010).

For researchers and practitioners there are some limitations, within the existing literature, and some promising approaches, applied primarily outside the social sciences, for reliably estimating growth. PM performance is limited to medians or averages for static performance and slope for growth over time (Foegen, Jiban, & Deno, 2007; Fuchs & Fuchs, 1998; Shinn, 1989; Wayman, Wallace, Wiley, Ticha, & Espin, 2007). The ordinary least squares (OLS) model is considered the most robust analytic model and present in some of the PM literature but results are primarily limited to slope (Deno, Fuchs, Marston, & Shin, 2001; Fogen, et al., 2007; Shinn, Good, & Stein, 1989; Wayman, et al., 2007). Christ and Coolong-Chaffin (2007) and Parker et al. (2010) identify other indices useful for interpreting results needed for important decisions. Methods new to education but applied in health, business and industry are emerging within the PM and single case research (Conover, 1980; Draper & Smith, 1998; Neter, Kutner, Nachtsheim, & Wasserman, 1996; Payton, Greenstone, & Schenker, 2003; Schenker & Gentleman, 2001; Sprent, 1993).

To contribute to the literature, this dissertation was a two-part study, based on texts outside the social sciences, to more reliably and precisely assess growth in PM results (Conover, 1980; Draper & Smith, 1998; Djulbegovic, Schwartz, & McMasters, 1999; Neter, et al., 1996; Schenker & Gentleman, 2001; Payton, Greenstone, & Schenker, 2003; Schenker & Gentleman, 2001; Sprent, 1993). The study examined results of 92 “low performing” third graders on the AIMSweb Mathematics Concepts and Applications (M-CAP) PM measure (NCS Pearson, 2009). The first part of the study applied the OLS model with a technique known as “approximate repeats” (Draper & Smith, 1998) to determine how much time was needed to reliably measure growth between point estimates of static performance. The approximate repeats method indicated that reliable growth occurred for 95% of the participants’ results between 10 and 16 weeks of progress monitoring by examining the presence of overlapping confidence intervals (Payton, et al., 2003; Schenker & Gentleman, 2001). The second part of the study compared the OLS model with non-parametric analyses (Conover, 1980; Sprent, 1993) to determine if there were differences between the models in terms of slopes, standard errors and other indices needed to judge if real growth occurred. Among the 92 “low performing” third graders’ PM results on the M-CAP, both models produced similar slopes and trendedness indices. The result summaries include trendedness indices to indicate stability of the slopes, standard error and confidence intervals to express precision of the trend, and a ratio of slope to SEslope to indicate which model had more power. While the OLS model had more power in 75% of the cases, 70% of the cases violated at least one parametric assumption. The results of this

study suggest that the nonparametric technique is a more favorable option for analyzing PM results. The results of this study have implications for practitioners, PM systems available to practitioners, and future research.

Implications for Practitioners

Assuming the results of this study can be replicated and vetted in the research there are many advantages for practitioners in terms of evaluating growth quickly and accurately. If practitioners can reliably estimate growth in 10 to 15 weeks, similar to the results found in this study, decision-making in terms of services and supports can be made more quickly. Decision-making using the indices in this study (i.e., Y_{est} , SEs, CIs) as indicated in the results of this study provided a precise estimate of static performance that included error. The standard error is most useful for practitioners to set confidence levels around point estimates or slopes at a level appropriate for the decision being considered. In this study CIs were set at 83.4% around point estimates so that when growth between two point estimates were compared of CIs did not overlap one could be 95% certain that real growth occurred (Payton, et al., 2003; Schenker & Gentleman, 2001). This provides a simple visual technique for examining growth between two points. Confidence bands were also set around slopes to precisely indicate at a level appropriate to the decision (i.e., in this study 95%) the range in which the true slope is likely to reside.

This study also extended PM literature in terms of analyzing slopes. This study compared the parametric OLS model with non-parametric analyses. Both analyses produced similar slopes and standard errors, needed to precisely estimate growth, and

trendedness indices, necessary for indicating score variability. Similar to previous research the OLS model had more power as indicated by the ratio of slope to SE_{slope} , but more than two-thirds of the cases did not meet parametric assumptions under the OLS model. This suggested that the nonparametric model provided more favorable results for analyzing and summarizing PM data. Currently practitioners do not have options for analyzing results using nonparametric models or results that include and error or trendedness indices.

The results of this study hold promise for practitioners in terms of summarizing results for important decisions, but the study also presents a new set of skills needed by practitioners. This dissertation does not address certain critical questions likely to emerge from the field. For example, is it reasonable to expect school districts to have staff capable of this degree of analysis? Who should have this training and what tools are needed to support school districts in analyzing results at this level? If PM results are used for high-stakes decisions will teacher evaluation systems judge teacher quality based on their PM results? These are all questions that need to be addressed by scholars and developers of PM systems.

Implications for Progress Monitoring Systems

If the field continues to argue that PM results should be used for important decisions than PM systems, used widely among practitioners, could support this change in practice. PM systems offer a number of useful data analysis tools that are readily accessible and meaningful to practitioners. These tools often include reports that provide individual student results such as graphic displays of performance with slope and goal

line, percentile rankings, comparisons to normed peer results, and programming recommendations—all to support decision-making. For medium- to high-stakes results examined in this study, practitioners need an indication of the stability of the slope from indices of trend, an estimate of performance level (the Y_{est}), the standard error of the Y_{est} and slope, and the capacity to set confidence limits befitting the level of decision under consideration. Practitioners would also need options to enter multiple data points in one week either by entering results from “multiple trials” or “approximate repeats” methods. Further it could be argued, based on the results of this study, that PM results be analyzed using nonparametric analyses rather than the typically applied OLS model. The system would need training components to support practitioners in understanding the benefits and weaknesses of those approaches as well as supporting their understanding of how to collect, analyze, and summarize results. Without appropriate support practitioners would be unlikely to see the value of these results in their decision-making. Progress monitoring system developers would also need to consider the feasibility of making changes to their system in terms of ease of development and cost to consumers. The changes identified here suggest significant changes to PM systems that are unlikely to occur without being valued by consumers or supported in research.

Implications for Future Research

Most importantly, more research is warranted before the methods described in this study can be introduced to practitioners or seriously considered by companies that offer PM systems. Only recently have researchers suggested that PM be used for important decisions (Christ & Coolong-Chaffin, 2007; Foegen et al., 2007; Wayman, et

al., 2007). The technical adequacy of static performance (Stage 1 PM research) and slope (Stage 2 PM research) have primarily been considered for low-stakes decisions (Fuchs, 2004). Issues with measurement error in the PM literature have been limited to studies of oral reading fluency PM results (Derr & Shapiro, 1989; Derr-Minneci & Shapiro, 1992; Francis, Santi, Barr, Fletcher, Varisco, & Foorman, 2008; Hintze & Christ, 2004; Hintze, Daly & Shapiro, 1998; Hintze, Owen, Shapiro, & Daly, 2000; Jenkins, Graft, & Miglioretti, 2009; Poncy, Skinner, & Axtell, 2005). Currently, the research is limited to summaries of slope, but slope as indicated in this study was insufficient for reporting growth on the M-CAP. Other than the results of this dissertation, methods for accounting for measurement error have not been studied using mathematics PM measures. More studies applying analytic techniques that account for error and provide precise and reliable summaries of performance are needed in the PM literature in order to seriously consider PM results for medium- to high-stakes decisions.

More studies are also needed that can address the limitations of the study in this dissertation. This dissertation used an extant dataset which lacked demographic information needed to generalize the results beyond this study. The researcher did not have basic demographic information such as gender, ethnicity, or socio-economic factors to determine if the dataset was representative of the population. The dataset did include numeric codes representing different districts, schools and classrooms, but specific information from which the sample was drawn (e.g., regions of the United States) were not provided to the researcher. Future research should include detailed information to

adequately assess threats to validity and possible ways in which the results would generalize to other samples.

Overall despite decades of research on PM, this phase of research devoted to using results for important decisions is still in its infancy. Researchers and practitioners should continue to support the field in identifying 1) appropriate methods that reliably and precisely estimate growth; 2) systems to analyze and summarize results that are accessible to practitioners; and 3) the potential misuse of this information. Currently, the use of PM for low-stakes decisions (e.g., screening or in-class decisions) is supported by the literature. The literature has not adequately assessed the appropriateness of these measures for medium- or high-stakes decisions.

REFERENCES

- Ardoin, S.P. & Christ, T.J. (2009) Curriculum based measurement of oral reading: Estimates of standard error when monitoring progress using alternate passage sets. *School Psychology Review*, 38, 266-283.
- Binder, C. (1990). Precision teaching and curriculum based measurement. *Journal of Precision Teaching*, 7, 33-35.
- Carnine, D. (2003). IDEA: Focusing on improvement results for children with disabilities. United States House of Representatives: March 13, 2003 testimony in the hearing before the subcommittee on Education Reform Committee on Education and the Workforce. Retrieved January 21, 2011 from <http://archives.republicans.edlabor.house.gov/archive/hearings/108th/edr/idea031303/carnine.htm>
- Case, L.P., Speece, D.L., & Molloy, D.E. (2003). The validity of a response-to-instruction paradigm to identify reading disabilities: A longitudinal analysis of individual differences and contextual factors. *School Psychology Review*, 32, 557-582.
- Christ, T.J. (2006). Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. *School Psychology Review*, 15, 128-133.
- Christ, T. J. & Coolong-Chaffin, M. (2007). Interpretations of curriculum-based measurement outcomes: Standard error and confidence intervals. *School Psychology Forum*, 1, 75-86.

- Clark, B. & Shinn, M.R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*, 234-248.
- Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral science* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates
- Conover, W.J. (1980). *Practical nonparametric statistics* (2nd ed.). New York, NY: John Wiley & Sons.
- Deno, S.L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.
- Deno, S.L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*, 184-192.
- 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.
- Derr, T.F. & Shapiro, E.S. (1989). A behavioral evaluation of curriculum-based assessment of reading. *Journal of Psychoeducational Assessment, 7*, 148-160.
- Derr-Minneci & Shapiro (1992). Validating curriculum-based measurement in reading from a behavioral perspective. *School Psychology Quarterly, 7*, 2-16.
- Djulgovic, B., Hozo, I., Schwartz, A., & McMasters, K.M. (1999). Acceptable regret in medical decision making. *Medical Hypothesis, 53*, 253-259.
- Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis*, (3rd ed.). New York, NY: John Wiley & Sons.

- Dunn, E.K. & Eckert, T.L. (2002). Curriculum-based measurement in reading: A comparison of similar versus challenging material. *School Psychology Quarterly*, *17*, 24-46.
- Foegen, A., Jiban, C., & Deno S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*, 121-139.
- Francis, D.J., Santi, K.L., Barr, C., Fletcher, J.M., Varisco, A. & Foorman, B.R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology*, *46*, 315-342.
- Franklin, R.D., Allison, D.B. & Gorman, B.S. (1997). *Design and analysis of single-case research*. Mahwah, NJ: Lawrence Erlbaum & Associates.
- Fuchs, L.S. (1986). Monitoring progress among mildly handicapped pupils: Review of current practice and research. *Remedial and Special Education*, *75*, 5-12.
- Fuchs, L.S. (2004). The past, present and future of curriculum-based measurement research. *School Psychology Review*, *33*, 188-192.
- Fuchs, L.S. & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice*, *13*, 204-219.
- Fuchs, L.S. Fuchs, D., Compton, D.L., Bryant, J.D., Hamlett, C.L. & Seethaler, P.M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children*, *73*, 311-330.
- Fuchs, L.S., Fuchs, D., Hamlett, C.L., Thompson, A., Roberts, P.H., Kupek, P., &

- Stecker, P. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Diagnostique, 19*, 23-49.
- Fuchs, L.S., Fuchs, D., Hamlett, C. L., Walz, L., Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*, 27-48.
- Fuchs, L.S., Fuchs, D., & Stecker, P.M. (1989). Effects of curriculum-based measurement on teachers' instructional planning. *Journal of Learning Disabilities, 22*, 51-59.
- Fuchs, L.S., Hamlett, C.L., & Fuchs, D. (1999). *Monitoring basic skills progress: Basic math concepts and applications* [computer program manual]. Austin, TX: PRO-ED, Inc.
- Good, R.H. & Shinn, M.R. (1990). Forecasting accuracy of slope estimates for reading curriculum-based measurement: Empirical evidence. *Behavioral Assessment, 12*, 179-193.
- Gorman, B.S. & Allison, D.B. (1996). Statistical alternatives for single-case designs. In Franklin, R.D., Allison, D.B., & Gorman, B.S. (Ed.), *Design and analysis of single-case research* (pp.159-214). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.
- Hambleton, R.K. & Jones, R.W. (1993). Comparison of classic test theory and item reponse theory and their application to test development. *Educational Measurement Issues and Practice, 12*, 38-47.
- Helwig, R. & Tindel, G. (2002). Using general outcome measures in mathematics to

- measure adequate yearly progress as mandated by Title I. *Assessment for Effective Intervention*, 28, 9-18.
- Hintze, J.M. & Christ, T.J. (2004). An examination of variability as a function of passage variance in CBM progress monitoring. *School Psychology Review*, 2, 204-217.
- Hintze, J.M., Daly III, E.J., & Shapiro, E.S. (1998). An investigation of the effects of passage difficulty level on outcomes of oral reading fluency progress monitoring. *School Psychology Review*, 27, 433-445.
- 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.
- Hollander, M. & Wolfe, D.A. (1999). *Nonparametric statistical methods* (2nd ed.). New York, NY: John Wiley & Sons, Inc.
- Individuals with Disabilities Education Improvement Act, Pub. L. 108-446 (2004).
- Jenkins, J.R., Graff, J.J., & Miglioretti, D.L. (2009). Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children*, 75, 151-163.
- Kazdin, A.E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.
- Manolov, R., Solanas, A., Sierra, V., & Evans, J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy*, 42, 533-545.
- Marston, D.B. (1989). A curriculum-based measurement approach to assessing academic

performance: What it is and why we do it. In Shinn, M.R. *Curriculum-based measurement: Assessing special children* (p. 18-78) New York, NY: The Guilford Press.

McGlinchey, M.T. & Hixson, M.D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review*, 33, 193-203.

National Council of Teachers of Mathematics (1980). *An agenda for action: Recommendations for school mathematics of the 1980s*. Reston, VA: Author

National Council of Teachers of Mathematics (2006). *Curriculum focal points for pre-kindergarten through grade 8: A quest for coherence*. Reston, VA: Author.

NCS Pearson (2009). AIMSweb mathematics concepts and applications administration and technical manual. Retrieved from <http://www.aimsweb.com/uploads/M-CAP%20Manual.pdf>

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. (1996). *Applied Linear Statistical Models* (4th ed.). Boston, MA: WCB McGraw-Hill.

No Child Left Behind Act of 2001, 20 U.S.C. §6301 et. seq.

Nunnally, J.C. (1967). *Psychometric theory*. New York, NY: McGraw-Hill Book Company.

Parker, R.I. & Brossart, D.F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy*, 34, 198-211.

Parker, R.I., Vannest, K.J., & Davis, J.L. (2011). Effect size in single case research: A review of nine nonoverlap techniques. *Behavior Modification*, 35, 303-322.

- Parker, R.I., Vannest, K.J., Davis, J.L., & Clemens, N.H. (2010). Defensible progress monitoring data for medium- and high-stakes decisions. *The Journal of Special Education*. Advance online publication. doi: 10.1177/0022466910376837
- Parker, R.I., Vannest, K.J., Davis, J.L. & Sauber, S.B. (2011). Combining non-overlapping and trend for single-case research: Tau-U. *Behavior Therapy*, 42, 284-299.
- Payton, M.E., Greenstone, M.H., & Schenker, N. (2003). Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science*, 3, 1-6.
- Pennypacker, H.S., Gutierrez, A., & Linsley, O.R. (2003). *Handbook of the standard acceleration chart*. Concord, MA: Cambridge Center for Behavioral Studies.
- Poncy, B.C., Skinner, C.H. & Axtel, P.K (2005). An investigation of the reliability and standard error of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment*, 23, 326-338.
- Restori, A.F., Gresham, F.M., & Cook, C.R. (2008). "Old habits die hard:" Past and current issues pertaining to response-to-intervention. *The California School Psychologist*, 13, 67-78.
- Salvia, J. & Ysseldyke, J.E. (2003). *Assessment in special and inclusive education*, (9th ed.). Boston, MA: Houghton Mifflin Company.
- Schatschneider, C., Wagner, R.K., & Crawford, E.C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences*, 18, 308-315.

- Schenker, N. & Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55, 182.
- Shapiro, E.S., Edwards, L., & Zigmond, N. (2005). Progress monitoring of mathematics among students with learning disabilities. *Assessment for Effective Intervention*, 30, 15-32.
- Shinn, M.R. (1989). Case study of Ann H: From referral to annual review. In Shinn, M.R. *Curriculum-based measurement: Assessing special children* (p. 79-89) New York, NY: The Guilford Press.
- Shinn, M.R., Good, R.H., & Stein, S. (1989). Summarizing trend in student achievement: A comparison of methods. *School Psychology Review*, 18, 356-370.
- Silberglitt, B. & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*, 23, 304-325.
- Speece, D.L. & Case, L.P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology*, 93, 735.
- Sprent, P. (1993). *Applied nonparametric statistical methods*, (2nd ed.). New York, NY: Chapman & Hall.
- Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York, NY: The Guilford Press.
- U.S. Department of Education Office of Special Education and Rehabilitative Services,

A New Era: Revitalizing Special Education for Children and Their Families,
Washington, DC, 2002

Wayman, M.M., Wallace, T., Wiley, H.I., Ticha, R. & Espin, C.A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education*, 41, 85-120.

Weiss, J. (2003 November 5). Rank-based nonparametric regression: Theil-sen estimator. Retrieved from
<http://www.unc.edu/courses/2003fall/biol145/001/docs/lectures/Oct27.html>

APPENDIX A

“LOW PERFORMING” PARTICIPANTS’ M-CAP PROGRESS MONITORING

RESULTS FOR PART 1 OF THE STUDY

TABLE 3.1. Participant Slope and Trendedness Indices on Third Grade M-CAPs

Student ID	Slope	R2	SEslope	Student ID	Slope	R2	SEslope
8597879	0.19	0.30	0.07	35382916	0.12	0.69	0.02
9840591	0.08	0.06	0.07	35696855	0.09	0.23	0.04
9840599	0.20	0.40	0.06	36923906	0.32	0.53	0.07
9840636	0.45	0.68	0.07	36924182	0.38	0.49	0.09
10738098	0.19	0.20	0.09	38363812	0.01	0.00	0.04
10739191	0.33	0.43	0.09	38363813	0.18	0.40	0.05
18062851	0.13	0.25	0.05	38363924	0.32	0.26	0.13
18337356	0.13	0.16	0.07	38363926	0.23	0.43	0.06
18822785	0.78	0.65	0.13	38363928	-0.01	0.01	0.04
19311652	0.40	0.62	0.07	38363929	0.15	0.10	0.10
19485419	0.29	0.58	0.06	38363965	0.46	0.67	0.07
19485494	0.22	0.35	0.07	38364845	0.27	0.73	0.04
19485530	0.32	0.50	0.07	38929384	0.60	0.83	0.06
19485963	0.44	0.37	0.13	39024705	0.16	0.25	0.06
23961789	0.33	0.62	0.06	43889271	0.42	0.52	0.09
26542501	0.55	0.68	0.09	43957949	0.41	0.91	0.03
27523261	0.46	0.73	0.06	43969748	0.29	0.50	0.07
28891518	0.18	0.13	0.11	44162379	0.09	0.04	0.10
31333552	0.27	0.44	0.07	44299660	0.22	0.66	0.04
31333553	0.60	0.82	0.07	44922885	0.41	0.69	0.06
31982440	0.69	0.79	0.08	45272220	0.30	0.45	0.08
32085739	0.41	0.61	0.07	45272223	0.28	0.41	0.08
32367075	0.10	0.29	0.04	45272224	0.11	0.12	0.07
32837773	0.26	0.29	0.10	45272227	0.23	0.17	0.12
32921873	0.43	0.56	0.09	45272229	0.87	0.83	0.09
33058186	0.49	0.64	0.08	45272237	0.39	0.63	0.07
33067964	0.97	0.76	0.12	45272238	0.15	0.40	0.04
33067965	0.48	0.78	0.06	45272270	0.14	0.20	0.06
33067969	0.42	0.48	0.10	45272272	0.25	0.50	0.06
33662217	0.71	0.63	0.13	45272275	0.17	0.29	0.06
33730300	0.46	0.66	0.08	45272276	0.02	0.02	0.04

TABLE 3.1. continued

Student ID	Slope	R2	SEslope	Student			
				ID	Slope	R2	SEslope
34429246	0.53	0.56	0.11	45272277	0.39	0.52	0.09
34433336	0.26	0.40	0.07	45272278	0.57	0.54	0.12
34433338	0.28	0.34	0.09	45272279	0.10	0.06	0.09
34433393	0.54	0.58	0.11	45272284	0.11	0.20	0.05
34440322	0.37	0.54	0.08	45272285	0.38	0.33	0.13
34440355	0.21	0.18	0.10	45272286	0.17	0.27	0.06
35375408	0.04	0.04	0.05	45272380	0.35	0.53	0.08
45272381	0.50	0.62	0.09	45272398	0.51	0.31	0.17
45272382	0.54	0.64	0.09	45272399	0.89	0.63	0.16
45272385	0.45	0.53	0.10	45272400	0.78	0.67	0.13
45272388	0.22	0.35	0.07	45272401	1.26	0.79	0.15
45272390	0.61	0.43	0.16	45454604	1.03	0.83	0.11
45272391	0.55	0.79	0.07	45456176	1.23	0.71	0.18
45272394	0.48	0.26	0.18	45724743	0.53	0.70	0.08
45272396	0.49	0.43	0.13	46195407	0.57	0.42	0.15
Avg	0.38	0.46	0.08				
Median	0.43	0.48	0.07				
SD	0.25	0.23	0.03				

Note. R²= Pearson R squared; SEslope= standard error of slope; Avg= average slope of the sample; Median= median of the sample; SD= standard deviation of the sample.

TABLE 3.2. Summaries of Improvement Using Approximate Repeats Method

Student	Yest & 83.4% CI @ wk 3	SE	Yest & 83.4% CI @ wk 10	SE	Yest & 83.4% CI @ wk 20	SE	Sign Growth @ wk 10 Y/N	Sign Growth @ wk 20 Y/N
8597879	3.29<4.27> 5.25	0.68	5.02<5.62>6.22	0.42	6.48 <7.54>8.60	0.74	N	Y
9840591	2.26<3.25>4.24	0.69	3.18<3.78>4.38	0.42	3.47<4.54>5.61	0.74	N	N
9840599	2.72<3.55> 4.38	0.58	4.44 <4.94>5.45	0.35	6.03 <6.93>7.83	0.63	Y	Y
9840636	1.83<2.87> 3.91	0.72	5.35 <5.98>6.62	0.44	9.31 <10.43>11.56	0.78	Y	Y
10738098	3.70<5.00> 6.29	0.90	5.62<6.42>7.21	0.55	7.04 <8.45>9.85	0.97	N	Y
10739191	0.19<1.45> 2.71	0.87	3.04 <3.81>4.57	0.53	5.81 <7.17>8.53	0.94	Y	Y
18062851	0.78<1.52> 2.26	0.52	2.03<2.48>2.94	0.32	3.05 <3.86>4.66	0.56	N	Y
18337356	4.43<5.41> 6.39	0.68	5.70<6.30>6.90	0.41	6.51 <7.57>8.63	0.74	N	Y
18822785	2.01<3.89> 5.77	1.30	8.21 <9.36>10.51	0.80	15.14 <17.18>19.21	1.41	Y	Y
19311652	2.52<3.59> 4.67	0.74	5.75 <6.41>7.06	0.45	9.27 <10.43>11.58	0.80	Y	Y
19485419	2.01<2.86> 3.70	0.59	4.38 <4.90>5.42	0.36	6.90 <7.82>8.73	0.64	Y	Y
19485494	4.05<5.07>6.09	0.71	6.01<6.63>7.25	0.43	7.77<8.87>9.97	0.76	N	Y
19485530	0.15<1.24> 2.34	0.76	2.74 <3.41>4.07	0.46	5.31 <6.49>7.68	0.82	Y	Y
19485963	5.84<7.76> 9.67	1.33	9.59<10.76>11.93	0.81	12.98 <15.05>17.13	1.44	N	Y
23961789	4.43<5.30> 6.17	0.60	7.05 <7.58>8.11	0.37	9.90 <10.84>11.78	0.65	Y	Y
26542501	2.64<3.96> 5.28	0.92	6.86 <7.66>8.47	0.56	11.52 <12.95>14.38	0.99	Y	Y
27523261	3.01<3.91> 4.82	0.63	6.60 <7.16>7.71	0.38	10.81 <11.79>12.77	0.68	Y	Y
28891518	4.23<5.80>7.37	1.09	5.98<6.93>7.89	0.66	6.86<8.56>10.25	1.18	N	N
31333552	2.43<3.44> 4.44	0.70	4.69 <5.30>5.92	0.43	6.89 <7.97>9.06	0.76	Y	Y
31333553	3.25<4.19> 5.13	0.65	7.87 <8.44>9.01	0.40	13.50 <14.51>15.53	0.70	Y	Y
31982440	2.67<3.85> 5.04	0.82	7.97 <8.69>9.41	0.50	14.32 <15.60>16.88	0.89	Y	Y
32085739	2.21<3.28> 4.36	0.74	5.51 <6.16>6.81	0.45	9.11 <10.27>11.43	0.80	Y	Y

TABLE 3.2. continued

Student	Yest & 83.4% CI @ wk 3	SE	Yest & 83.4% CI @ wk 10	SE	Yest & 83.4% CI @ wk 20	SE	Sign Growth @ wk 10 Y/N	Sign Growth @ wk 20 Y/N
32367075	1.20<1.73>2.26	0.37	2.10<2.42>2.75	0.23	2.84<3.42>3.99	0.40	N	Y
32837773	5.62<7.02>8.42	0.97	7.98<8.84>9.69	0.59	9.91<11.43>12.94	1.05	N	Y
32921873	4.24<5.55>6.87	0.91	7.69<8.49>9.29	0.56	11.26<12.67>14.09	0.98	Y	Y
33058186	1.43<2.66>3.89	0.85	5.29<6.04>6.79	0.52	9.54<10.87>12.20	0.92	Y	Y
33067964	3.46<5.38>7.29	1.33	10.88<12.05>13.22	0.81	19.50<21.58>23.65	1.44	Y	Y
33067965	4.55<5.45>6.35	0.62	8.22<8.76>9.31	0.38	12.53<13.50>14.47	0.67	Y	Y
33067969	2.95<4.38>5.81	1.00	6.46<7.34>8.21	0.61	10.02<11.57>13.12	1.08	Y	Y
33662217	5.68<7.52>9.36	1.28	11.32<12.44>13.56	0.78	17.48<19.47>21.46	1.38	Y	Y
33730300	2.62<3.76>4.90	0.79	6.23<6.93>7.62	0.48	10.23<11.46>12.69	0.85	Y	Y
34429246	4.92<6.46>8.01	1.07	9.24<10.18>11.13	0.65	13.83<15.50>17.17	1.16	Y	Y
34433336	2.53<3.61>4.70	0.75	4.75<5.41>6.07	0.46	6.80<7.98>9.15	0.82	Y	Y
34433338	4.98<6.26>7.54	0.89	7.42<8.20>8.98	0.54	9.59<10.97>12.36	0.96	N	Y
34433393	2.48<4.01>5.55	1.06	6.90<7.84>8.77	0.65	11.64<13.29>14.95	1.15	Y	Y
34440322	2.17<3.34>4.52	0.82	5.12<5.83>6.55	0.50	8.12<9.39>10.67	0.88	Y	Y
34440355	3.30<4.75>6.20	1.01	5.25<6.14>7.02	0.61	6.54<8.11>9.68	1.09	N	Y
35375408	2.24<2.99>3.74	0.52	2.83<3.29>3.75	0.32	2.91<3.72>4.53	0.56	N	N
35382916	1.54<1.83>2.11	0.20	2.51<2.69>2.86	0.12	3.61<3.92>4.23	0.21	Y	Y
35696855	3.11<3.63>4.15	0.36	3.93<4.25>4.56	0.22	4.56<5.12>5.68	0.39	N	Y
36923906	3.28<4.29>5.31	0.71	5.92<6.54>7.16	0.43	8.64<9.74>10.84	0.76	Y	Y
36924182	2.57<3.87>5.18	0.90	5.69<6.48>7.28	0.55	8.81<10.21>11.62	0.98	Y	Y
38363812	3.47<4.11>4.75	0.44	3.75<4.14>4.53	0.27	3.49<4.18>4.87	0.48	N	N
38363813	1.82<2.54>3.27	0.50	3.38<3.82>4.26	0.31	4.85<5.64>6.42	0.54	Y	Y
38363924	2.46<4.29>6.12	1.27	5.42<6.54>7.65	0.78	7.76<9.74>11.72	1.38	N	Y
38363926	3.74<4.65>5.56	0.63	5.65<6.21>6.76	0.38	7.45<8.43>9.41	0.68	Y	Y

TABLE 3.2. continued

Student	Yest & 83.4% CI @ wk 3	SE	Yest & 83.4% CI @ wk 10	SE	Yest & 83.4% CI @ wk 20	SE	Sign Growth @ wk 10 Y/N	Sign Growth @ wk 20 Y/N
38363928	2.35<2.95>3.54	0.42	2.46<2.83>3.19	0.25	2.01<2.66>3.30	0.45	N	N
38363929	2.41<3.93>5.44	1.05	3.98<4.91>5.83	0.64	4.67<6.31>7.95	1.14	N	N
38363965	1.99<3.05> 4.11	0.73	5.61 <6.26>6.90	0.45	9.69 <10.84>11.98	0.79	Y	Y
38364845	1.29<1.82> 2.35	0.37	3.41 <3.73>4.05	0.22	5.88 <6.45>7.02	0.40	Y	Y
38929384	3.63<4.61> 5.59	0.68	8.15 <8.74>9.34	0.41	13.59 <14.65>15.70	0.73	Y	Y
39024705	4.61<5.54> 6.46	0.64	6.13<6.69>7.26	0.39	7.34 <8.34>9.34	0.70	N	Y
43889271	3.63<5.00> 6.37	0.95	7.04 <7.87>8.71	0.58	10.50 <11.98>13.47	1.03	Y	Y
43957949	3.10<3.54> 3.99	0.31	6.13 <6.40>6.67	0.19	10.00 <10.48>10.97	0.33	Y	Y
43969748	3.25<4.25> 5.25	0.69	5.63 <6.24>6.85	0.42	8.00 <9.08>10.16	0.75	Y	Y
44162379	2.46<3.92>5.38	1.01	3.60<4.49>5.38	0.62	3.73<5.30>6.88	1.09	N	N
44299660	0.08<0.61> 1.14	0.37	1.84 <2.16>2.48	0.22	3.79 <4.37>4.94	0.40	Y	Y
44922885	1.84<2.75> 3.67	0.64	5.12 <5.68>6.24	0.39	8.86 <9.85>10.85	0.69	Y	Y
45272220	3.67<4.76> 5.84	0.76	6.18 <6.84>7.51	0.46	8.65 <9.83>11.01	0.82	Y	Y
45272223	2.80<3.96> 5.13	0.81	5.16 <5.87>6.58	0.49	7.33 <8.59>9.85	0.87	Y	Y
45272224	3.83<4.79>5.74	0.66	4.89<5.47>6.06	0.41	5.42<6.45>7.49	0.72	N	N
45272227	3.07<4.76> 6.44	1.17	5.36<6.39>7.41	0.71	6.90 <8.72>10.54	1.26	N	Y
45272229	2.93<4.32> 5.70	0.96	9.53<10.37>11.22	0.59	17.53 <19.02>20.52	1.04	Y	Y
45272237	2.33<3.35> 4.37	0.71	5.42 <6.04>6.66	0.43	8.79 <9.89>10.99	0.76	Y	Y
45272238	1.32<1.92> 2.51	0.41	2.63 <2.99>3.35	0.25	3.88 <4.52>5.16	0.45	Y	Y
45272270	3.31<4.21> 5.11	0.62	4.60<5.15>5.70	0.38	5.53 <6.50>7.47	0.68	N	Y
45272272	1.87<2.73> 3.59	0.60	3.90 <4.42>4.95	0.37	5.91 <6.85>7.78	0.65	Y	Y
45272275	1.82<2.70> 3.58	0.61	3.30<3.84>4.37	0.37	4.51 <5.46>6.41	0.66	N	Y
45272276	2.14<2.75>3.35	0.42	2.52<2.89>3.25	0.26	2.43<3.08>3.74	0.45	N	N
45272277	3.28<4.56> 5.83	0.88	6.50 <7.28>8.06	0.54	9.79 <11.17>12.54	0.96	Y	Y

TABLE 3.2. continued

Student	Yest & 83.4% CI @ wk 3	SE	Yest & 83.4% CI @ wk 10	SE	Yest & 83.4% CI @ wk 20	SE	Sign Growth @ wk 10 Y/N	Sign Growth @ wk 20 Y/N
45272278	5.28<7.04> 8.81	1.23	9.97 <11.05>12.12	0.75	14.86 <16.77>18.68	1.32	Y	Y
45272279	2.86<4.18>5.51	0.92	4.09<4.90>5.71	0.56	4.49<5.92>7.35	0.99	N	N
45272284	2.09<2.81> 3.53	0.50	3.04<3.48>3.92	0.31	3.65 <4.43>5.21	0.54	N	Y
45272285	4.59<6.42> 8.26	1.27	8.02<9.14>10.26	0.78	11.03 <13.01>14.99	1.38	N	Y
45272286	3.57<4.50> 5.43	0.65	5.08<5.65>6.22	0.40	6.27 <7.28>8.29	0.70	N	Y
45272380	2.04<3.14> 4.24	0.76	4.93 <5.60>6.27	0.47	7.93 <9.12>10.31	0.83	Y	Y
45272381	0.83<2.14> 3.45	0.91	4.84 <5.64>6.44	0.55	9.23 <10.64>12.06	0.98	Y	Y
45272382	1.65<3.08> 4.51	0.99	5.89 <6.76>7.63	0.61	10.47 <12.02>13.57	1.07	Y	Y
45272385	0.78<2.25> 3.72	1.02	4.47 <5.36>6.26	0.62	8.23 <9.81>11.40	1.10	Y	Y
45272388	3.14<4.15> 5.16	0.70	5.15<5.77>6.38	0.43	6.99 <8.09>9.18	0.76	N	Y
45272390	-1.41<1.00> 3.41	1.67	3.57 <5.04>6.51	1.02	8.21 <10.81>13.42	1.81	Y	Y
45272391	3.94<4.96> 5.97	0.71	8.17 <8.79>9.41	0.43	13.16 <14.26>15.36	0.76	Y	Y
45272394	-1.68<1.06> 3.79	1.90	2.50<4.17>5.84	1.16	5.67 <8.62>11.58	2.05	N	Y
45272396	4.39<6.30> 8.21	1.33	8.50 <9.66>10.83	0.81	12.40 <14.47>16.53	1.43	Y	Y
45272398	-0.85<1.74> 4.33	1.80	3.51<5.09>6.67	1.10	7.08 <9.88>12.68	1.95	N	Y
45272399	2.46<4.64> 6.83	1.52	9.70 <11.04>12.37	0.93	17.81 <20.17>22.54	1.64	Y	Y
45272400	2.34<4.28> 6.22	1.34	8.40 <9.58>10.76	0.82	15.05 <17.14>19.24	1.45	Y	Y
45272401	1.53<3.89> 6.26	1.64	11.13 <12.57>14.01	1.00	22.41 <24.97>27.52	1.77	Y	Y
45454604	2.93<4.55> 6.16	1.12	10.71 <11.69>12.67	0.68	20.16 <21.90>23.65	1.21	Y	Y
45456176	6.18<8.81> 11.43	1.82	15.79 <17.39>18.99	1.11	26.82 <29.66>32.49	1.97	Y	Y
45724743	3.65<4.79> 5.93	0.79	7.86 <8.56>9.25	0.48	12.70 <13.94>15.17	0.86	Y	Y
46195407	-1.99<.19> 2.38	1.52	2.98 <4.32>5.65	0.93	7.84 <10.21>12.57	1.64	Y	Y

Note. Y_{est} = estimated Y or Yhat; CI = confidence interval for the Y_{est} score preset at 83.4%; non-overlapping confidence bands in boldface; SE = standard error of the Y_{est} ; Sign Growth Y/N = significant growth at week 10 and 20, yes or no.

TABLE 3.3. Sensitivity to Growth Using Approximate Repeats Method

Student	Week Significant Growth Detected Between Weeks 10 and 20										
	Wk 10	Wk 11	Wk 12	Wk 13	Wk 14	Wk 15	Wk 16	Wk 17	Wk 18	Wk 19	Wk 20
8597879	N		Y								Y
10738098	N						Y				Y
18062851	N			Y							Y
18337356	N									Y	Y
19485494	N	Y									Y
19485963	N	Y									Y
32367075	N		Y								Y
32837773	N		Y								Y
34433338	N	Y									Y
34440355	N						Y				Y
35696855	N						Y				Y
38363924	N			Y							Y
39024705	N			Y							Y
45272227	N						Y				Y
45272270	N				Y						Y
45272275	N		Y								Y
45272284	N						Y				Y
45272285	N	Y									Y
45272286	N		Y								Y
45272388	N	Y									Y
45272394	N			Y							Y
45272398	N		Y								Y

Note. Wk = week; N= no, significant growth not detected; Y= yes, significant growth detected.

TABLE 3.4. Comparison of T-test Results to Overlapping Confidence Intervals Technique

Student	t-stat wk 3 & 10	P value	Y_{est} diff. between wk 3 & 10 at 95%CI	t-stat wk 3 & 20	P value	Y_{est} diff. between wk 3 & 20 at 95%CI	Sign Growth @ wk 10 Y/N	Sign Growth @ wk 20 Y/N	Same Results w/Overlap. CIs Method Y/N
8597879	1.69	0.10	-0.27<1.35>2.97	3.25	0.00	1.24<3.27>5.30	N	Y	Y
9840591	0.66	0.52	-1.10 <0.53>2.16	1.28	0.21	-0.75 <1.29>3.33	N	N	Y
9840599	2.05	0.00	1.40<3.11>4.82	7.12	0.00	5.41<7.56>9.71	Y	Y	Y
9840636	3.69	0.01	1.40<3.11>4.82	0.00	0.00	5.41<7.56>9.71	Y	Y	Y
10738098	1.35	0.19	-0.71<1.42>3.55	2.61	0.01	0.78<3.45>6.12	N	Y	Y
10739191	0.00	2.32	0.03 <2.36>4.42	4.47	0.00	3.13 <5.72>8.31	Y	Y	Y
18062851	0.00	1.57	0.27 <.96>2.19	3.06	0.00	0.80<2.34>3.88	N	Y	Y
18337356	1.12	0.27	0.71 <0.89>2.49	2.15	0.04	0.13<2.16>4.19	N	Y	Y
18822785	3.58	0.00	2.38 <5.47>8.56	6.93	0.00	9.41 <13.29>17.17	Y	Y	Y
19311652	3.26	0.00	1.07<2.82>4.57	6.28	0.00	4.64<6.84>9.04	Y	Y	Y
19485419	2.95	0.01	0.64 <2.04>3.44	5.70	0.00	3.20 <4.96>6.72	Y	Y	Y
19485494	1.88	0.07	-0.12 <1.56>3.24	3.65	0.00	1.70 <3.80>5.90	N	Y	Y
19485530	2.44	0.02	0.37<2.17>3.97	4.70	0.00	2.99<5.25>7.51	Y	Y	Y
19485963	1.93	0.06	-0.15 <3.00>6.15	3.72	0.00	3.33 <7.29>11.25	N	Y	Y
23961789	3.23	0.00	0.86 <2.28>3.70	6.26	0.00	3.75 <5.54>7.33	Y	Y	Y
26542501	3.44	0.00	1.52 <3.70>5.88	6.65	0.00	6.26<8.99>11.72	Y	Y	Y
27523261	4.42	0.00	1.70<3.25>4.74	8.50	0.00	6.01<7.88>9.75	Y	Y	Y
28891518	0.87	0.38	-1.45 <1.13>3.71	1.72	0.09	-0.49 <2.76>6.01	N	N	Y
31333552	2.26	0.03	0.20<1.86>3.52	4.38	0.00	2.44 <4.53>6.62	Y	Y	Y
31333553	5.57	0.00	2.71 <4.25>5.79	10.80	0.00	8.39 <10.32>12.25	Y	Y	Y
31982440	5.04	0.00	2.90 <4.84>6.78	9.71	0.00	9.30 <11.75>14.20	Y	Y	Y

TABLE 3.4. continued

Student	t-stat	P value	Y _{est} diff. between	t-stat	P value	Y _{est} diff. between	Sign	Sign	Same Results w/Overlap. CIs Method Y/N
	wk 3 & 10		wk 3 & 10 at 95%CI	wk 3 & 20		wk 3 & 20 at 95%CI	@ wk 10 Y/N	@ wk 20 Y/N	
32085739	3.33	0.00	1.13<2.88>4.63	6.41	0.00	4.79<6.99>9.14	Y	Y	Y
32367075	1.58	0.12	-0.19 <0.69>1.57	3.10	0.00	0.58<1.69>2.79	N	Y	Y
32837773	1.6	0.12	-0.47<1.82>4.11	2.75	0.01	0.78 <2.94>5.10	N	Y	Y
32921873	2.75	0.00	0.78 <2.94>5.10	5.32	0.00	4.42 <7.12>9.82	Y	Y	Y
33058186	3.39	0.00	1.36<3.38>5.39	6.55	0.00	5.68<8.21>10.74	Y	Y	Y
33067964	4.28	0.00	3.52<6.67>9.82	8.26	0.00	12.24<16.20>20.16	Y	Y	Y
33067965	2.97	0.01	1.19 <3.72>6.25	5.73	0.00	5.85 <9.04>12.23	Y	Y	Y
33067969	2.05	0.05	0.02<1.80>3.58	3.93	0.00	2.12<4.37>6.62	Y	Y	Y
33662217	3.28	0.00	1.89<4.92>7.95	6.35	0.00	8.15<11.95>15.75	Y	Y	Y
33730300	3.43	0.00	1.30<3.17>5.04	6.64	0.00	5.35<7.70>10.05	Y	Y	Y
34429246	2.97	0.01	1.19<3.72>6.25	5.73	0.00	5.58<9.04>12.23	Y	Y	Y
34433336	2.05	0.05	0.02<1.80>3.58	3.93	0.00	2.12<4.37>6.62	Y	Y	Y
34433338	1.86	0.07	-0.16<1.94>4.04	3.6	0.00	2.06<4.71>7.36	Y	Y	Y
34433393	3.08	0.00	1.32 <3.83>6.34	5.93	0.00	6.12 <9.28>12.44	Y	Y	Y
34440322	2.59	0.01	0.55 <2.49>4.43	5.03	0.00	3.62 <6.05>8.48	Y	Y	Y
34440355	1.18	0.25	-0.99<1.39>3.77	2.26	0.03	0.36<3.36>6.36	N	Y	Y
35375408	0.39	0.70	-1.24<0.30>1.84	0.96	0.35	-0.81<0.73>2.27	N	N	Y
35382916	3.69	0.00	0.39<0.86>1.33	7.21	0.00	1.50<2.09>2.68	Y	Y	Y
35696855	1.47	0.15	-0.23<0.62>1.47	2.81	0.01	0.42<1.49>2.56	N	Y	Y
36923906	2.71	0.01	0.57<2.25>3.93	5.24	0.00	3.35<5.45>7.55	Y	Y	Y
36924182	2.47	0.02	0.48<2.61>4.74	4.76	0.00	3.65<6.34>9.03	Y	Y	Y
38363812	0.06	0.95	-1.01<0.03>1.07	0.11	0.91	-1.25<0.07>1.39	N	N	Y

TABLE 3.4. continued

Student	t-stat	P value	Y _{est} diff. between	t-stat	P value	Y _{est} diff. between	Sign	Sign	Same Results w/Overlap. CIs Method Y/N
	wk 3 & 10		wk 3 & 10 at 95%CI	wk 3 & 20		wk 3 & 20 at 95%CI	@ wk 10 Y/N	@ wk 20 Y/N	
38363813	2.18	0.04	0.09<1.28>2.46	4.21	0.00	1.61<3.10>4.59	Y	Y	Y
38363924	1.51	0.14	-0.76<2.25>5.26	2.91	0.01	1.66<5.45>9.24	N	Y	Y
38363926	2.12	0.04	0.07<1.56>3.05	4.08	0.00	1.91<3.78>5.65	Y	Y	Y
38363928	0.25	0.81	-1.11<-0.12>0.87	0.47	0.64	-1.53<-0.21>0.95	N	N	Y
38363929	0.8	0.43	-1.51<0.98>3.47	1.54	0.13	-0.75<2.38>5.51	N	N	Y
38363965	3.74	0.00	1.48<3.21>4.94	7.24	0.00	5.62<7.79>9.96	Y	Y	Y
38364845	4.44	0.00	1.04<1.91>2.78	8.5	0.00	3.53<4.63>5.73	Y	Y	Y
38929384	5.2	0.00	2.53<4.13>5.73	10.06	0.00	8.02<10.04>12.06	Y	Y	Y
39024705	1.53	0.13	-0.36<1.15>2.66	2.95	0.01	0.88<2.80>4.72	N	Y	Y
43889271	2.58	0.01	0.62<2.87>5.12	4.98	0.00	4.15<6.98>9.18	Y	Y	Y
43957949	7.87	0.00	2.13<2.86>3.59	15.33	0.00	6.02<6.94>7.86	Y	Y	Y
43969748	2.46	0.02	0.36<1.99>3.62	4.74	0.00	2.77<4.83>6.89	Y	Y	Y
44162379	0.48	0.63	-1.83<0.57>2.97	0.93	0.36	-1.62<1.38>4.38	N	N	Y
44299660	3.6	0.00	0.68<1.55>2.42	6.9	0.00	3.66<3.76>4.86	Y	Y	Y
44922885	3.91	0.00	1.42<2.93>4.44	7.54	0.00	5.18<7.10>9.00	Y	Y	Y
45272220	2.34	0.02	0.28<2.08>3.88	4.53	0.00	2.81<5.07>7.33	Y	Y	Y
45272223	2.02	0.05	-0.03<1.91>3.82	3.9	0.00	2.23<4.63>7.03	N	Y	Y
45272224	0.88	0.39	-0.89<0.68>2.25	1.7	0.1	-0.31<1.66>3.63	N	N	Y
45272227	1.91	0.24	-1.14<1.36>4.40	2.3	0.03	0.48<3.96>7.44	N	Y	Y
45272229	5.37	0.00	3.77<6.05>8.33	10.39	0.00	11.84<14.70>17.56	Y	Y	Y
45272237	3.24	0.00	1.01<2.69>4.37	6.29	0.00	4.44<6.54>8.64	Y	Y	Y
45272238	2.23	0.03	0.10<1.07>2.04	4.27	0.00	1.37<2.60>3.83	Y	Y	Y

TABLE 3.4. continued

Student	t-stat	P value	Y _{est} diff. between	t-stat	P value	Y _{est} diff. between	Sign	Sign	Same Results w/Overlap. CIs Method Y/N
	wk 3 & 10		wk 3 & 10 at 95%CI	wk 3 & 20		wk 3 & 20 at 95%CI	@ wk 10 Y/N	@ wk 20 Y/N	
45272270	1.29	0.20	-0.53<0.94>2.41	2.49	0.02	0.43<2.29>4.15	N	Y	Y
45272272	2.4	0.02	0.27<1.69>3.11	4.66	0.00	2.33<4.12>5.91	Y	Y	Y
45272275	1.58	0.12	-0.30<1.14>2.58	3.07	0.00	0.94<2.76>4.58	N	Y	Y
45272276	0.28	0.78	-0.85<0.14>1.14	0.54	0.59	-0.91<0.33>1.57	N	N	Y
45272277	2.63	0.01	0.63<2.72>4.81	5.08	0.00	3.98<6.61>9.24	Y	Y	Y
45272278	2.78	0.01	1.10<4.01>6.92	5.39	0.00	6.08<9.73>13.37	Y	Y	Y
45272279	0.67	0.51	-1.46<0.72>2.90	1.29	0.21	-0.99<1.74>4.47	N	N	Y
45272284	1.14	0.26	-0.52<0.67>1.86	2.2	0.03	0.13<1.62>3.11	N	Y	Y
45272285	1.83	0.08	-0.29<2.72>5.73	3.51	0.00	2.80<6.59>10.38	N	Y	Y
45272286	1.51	0.14	-0.39<1.15>2.69	2.91	0.01	0.85<2.78>4.71	N	Y	Y
45272380	2.75	0.01	0.65<2.46>4.29	5.31	0.00	3.71<5.98>8.25	Y	Y	Y
45272381	3.29	0.00	1.35<3.50>5.65	6.36	0.00	5.80<8.50>11.20	Y	Y	Y
45272382	3.16	0.00	1.33<3.68>6.03	6.13	0.00	5.99<8.94>11.89	Y	Y	Y
45272385	2.61	0.01	0.70<3.11>5.52	5.04	0.00	4.53<7.56>10.59	Y	Y	Y
45272388	1.97	0.06	-0.04<1.62>3.28	3.81	0.00	1.85<3.94>6.02	N	Y	Y
45272390	2.06	0.05	0.09<4.04>8.00	3.98	0.00	4.83<9.81>14.79	Y	Y	Y
45272391	4.61	0.00	2.15<3.83>5.51	8.94	0.00	7.20<9.30>11.40	Y	Y	Y
45272394	1.4	0.17	-1.38<3.11>7.61	2.7	0.01	1.91<7.56>13.20	N	Y	Y
45272396	2.16	0.04	0.21<3.36>6.51	4.18	0.00	4.22<8.17>12.12	Y	Y	Y
45272398	1.59	0.12	-0.91<3.35>7.61	3.07	0.00	2.78<8.14>13.50	N	Y	Y
45272399	3.59	0.00	2.80<6.40>10.00	6.95	0.00	11.01<15.53>20.05	Y	Y	Y
45272400	3.37	0.00	2.12<5.30>8.48	6.51	0.00	8.87<12.86>16.85	Y	Y	Y

TABLE 3.4. continued

Student	t-stat wk 3 & 10	P value	Y _{est} diff. between wk 3 & 10 at 95%CI	t-stat wk 3 & 20	P value	Y _{est} diff. between wk 3 & 20 at 95%CI	Sign Growth @ wk 10 Y/N	Sign Growth @ wk 20 Y/N	Same Results w/Overlap. CIs Method Y/N
45272401	4.52	0.00	4.80<8.68>12.56	8.74	0.00	16.20<21.08>25.96	Y	Y	Y
45454604	5.45	0.00	4.49<7.14>9.79	10.52	0.00	14.02<17.35>20.68	Y	Y	Y
45456176	4.02	0.00	4.27<8.58>12.89	7.77	0.00	15.43<20.85>26.27	Y	Y	Y
45724743	4.08	0.00	1.90<3.77>5.64	7.86	0.00	6.82<9.18>11.54	Y	Y	Y
46195407	2.32	0.03	0.53<4.13>7.73	4.48	0.00	5.50<10.02>14.54	Y	Y	Y

Notes. t-stat @ wk 3 and 10 and week 3 and 20= t-test results comparing Yest at week 3 to week 10 and week 3 to week 20; P value= probability calculation from t-test; Sign Growth @ wk 10 Y/N and wk 20 Y/N= statistical significance tests indicate growth was significant at week 10, yes or no and week 20, yes or no. Same Results w/Overlap CIs Method Y/N= t-statistics matched visual technique of overlapping confidence intervals, yes or no.

APPENDIX B

“LOW PERFORMING” PARTICIPANTS’ M-CAP PROGRESS MONITORING RESULTS FOR PART 2 OF THE STUDY

TABLE 4.1 Descriptive Results for Slopes

Slopes	Range	Mean	Median	SD
LR Slope	-0.01 to 1.26	0.38	0.33	0.25
Theil-Sen Estimator	0.00 to 1.23	0.36	0.31	0.24

Note. Models= parametric and non-parametric analyses of slope; Range= range of slope size within participants sampled; Mean= average slope size; Median= middle slope value within participants sampled; SD= standard deviation of the slopes

TABLE 4.2. Individual Calculations of Slope Size, Trendedness and Standard Errors and Confidence Intervals

Student	<u>OLS Analysis</u>				<u>Tau Analysis</u>			
	LR slope	R2	SE slope	95% CI for R Coeff.	Theil -Sen	Tau-U	SE slope	95% CI for Theil-Sen
45272401	1.26	0.79	0.15	0.95<1.26>1.58	1.21	0.13	0.13	0.50<0.57>0.64
45456176	1.23	0.71	0.18	0.85<1.23>1.60	1.23	0.07	0.07	0.29<0.33>0.43
45454604	1.03	0.83	0.11	0.81<1.03>1.25	1.00	0.29	0.29	1.13<1.23>1.47
33067964	<u>0.97</u>	0.76	0.12	0.72<0.97>1.23	<u>0.44</u>	0.50	0.05	0.13<0.15>0.19
45272399	0.89	0.63	0.16	0.56<0.89>1.22	0.75	0.25	0.25	1.07<1.21>1.29
45272229	0.87	0.83	0.09	0.69<0.87>1.06	0.85	0.20	0.20	0.93<1.00>1.05
18822785	0.78	0.65	0.13	0.50<0.78>1.05	0.71	0.07	0.07	0.20<0.25>0.30
45272400	0.78	0.67	0.13	0.52<0.78>1.04	0.75	0.19	0.19	0.63<0.750>0.80
33662217	0.71	0.63	0.13	0.44<0.71>0.97	0.67	0.16	0.16	0.41<0.50>0.63

TABLE 4.2. continued

Student	<u>OLS Analysis</u>				<u>Tau Analysis</u>			
	LR slope	R2	SE slope	95% CI for R Coeff.	Theil -Sen	Tau-U	SE slope	95% CI for Theil-Sen
31982440	0.69	0.79	0.08	0.52<0.69>0.86	0.73	0.17	0.17	0.67<0.75>0.83
45272390	<u>0.61</u>	0.43	0.16	0.27<0.61>0.94	<u>0.38</u>	0.51	0.13	0.33<0.42>0.50
31333553	0.60	0.82	0.07	0.46<0.60>0.74	0.51	0.11	0.11	0.41<0.47>0.53
38929384	0.60	0.83	0.06	0.47<0.60>0.73	0.61	0.07	0.07	0.13<0.20>0.33
46195407	<u>0.57</u>	0.42	0.15	0.25<0.57>0.90	<u>0.33</u>	0.63	0.12	0.40<0.50>0.53
45272278	0.57	0.54	0.12	0.32<0.57>0.83	0.59	0.13	0.13	0.50<0.55>0.60
46542501	0.55	0.68	0.09	0.37<0.55>0.72	0.50	0.13	0.13	0.40<0.50>0.60
45272391	0.55	0.79	0.07	0.42<0.55>0.69	0.55	0.10	0.10	0.14<0.25>0.33
45272382	0.54	0.64	0.09	0.35<0.54>0.74	0.47	0.13	0.13	0.32<0.41>0.50
34433393	0.54	0.58	0.11	0.32<0.54>0.76	0.54	0.09	0.09	0.27<0.33>0.41
34429246	0.53	0.56	0.11	0.30<0.53>0.75	0.50	0.11	0.11	0.29<0.37>0.43
45724743	0.53	0.70	0.08	0.37<0.53>0.70	0.57	0.10	0.10	0.33<0.38>0.50
45272398	<u>0.51</u>	0.31	0.17	0.14<0.51>0.87	<u>0.25</u>	0.57	0.11	0.29<0.40>0.47
45272381	0.50	0.62	0.09	0.31<0.50>0.69	0.50	0.08	0.08	0.13<0.20>0.22
33058186	0.49	0.64	0.08	0.31<0.49>0.66	0.50	0.05	0.05	0.00<0.08>0.14
45272396	0.49	0.43	0.13	0.22<0.49>0.76	0.50	0.14	0.14	0.30<0.42>0.50
45272394	<u>0.48</u>	0.26	0.18	0.09<0.48>0.86	<u>0.20</u>	0.37	0.04	0.06<0.10>0.13
33067965	0.48	0.78	0.06	0.36<0.48>0.60	0.47	0.00	0.00	0.00<0.00>0.00
33730300	0.46	0.66	0.08	0.30<0.46>0.62	0.46	0.07	0.07	0.13<0.08>0.22
27523261	0.46	0.73	0.06	0.33<0.46>0.59	0.50	0.17	0.17	0.50<0.59>0.67
38363965	0.46	0.67	0.07	0.30<0.46>0.61	0.50	0.07	0.07	0.00<0.07>0.17
45272385	<u>0.45</u>	0.53	0.10	0.25<0.45>0.66	<u>0.33</u>	0.26	0.10	0.00<0.16>0.21
9840636	0.45	0.68	0.07	0.30<0.45>0.60	0.46	0.08	0.08	0.20<0.25>0.30
19485963	0.44	0.37	0.13	0.16<0.44>0.71	0.46	0.07	0.07	0.08<0.14>0.20

TABLE 4.2. continued

Student	<u>OLS Analysis</u>				<u>Tau Analysis</u>			
	LR slope	R2	SE slope	95% CI for R Coeff.	Theil -Sen	Tau-U	SE slope	95% CI for Theil-Sen
32921873	0.43	0.56	0.09	0.25<0.43>0.61	0.38	0.42	0.08	0.14<0.22>0.27
43889271	0.42	0.52	0.09	0.22<0.42>0.61	0.43	0.77	0.10	0.75<0.85>1.00
33067969	0.42	0.48	0.10	0.21<0.42>0.63	0.44	0.59	0.10	0.33<0.39>0.44
44922885	0.41	0.69	0.06	0.28<0.41>0.55	0.40	0.61	0.05	0.18<0.21>0.25
43957949	0.41	0.91	0.03	0.35<0.41>0.47	0.41	0.51	0.10	0.27<0.33>0.40
32085739	0.41	0.61	0.07	0.25<0.41>0.56	0.43	0.33	0.10	0.15<0.21>0.33
19311652	0.40	0.62	0.07	0.25<0.4>0.56	0.40	0.13	0.08	0.00<0.07>0.18
45272237	0.39	0.63	0.07	0.25<0.39>0.53	0.39	0.84	0.08	0.38<0.41.>0.44
45272277	0.39	0.52	0.09	0.21<0.39>0.57	0.41	0.62	0.10	0.35<0.40>0.50
36924182	0.38	0.49	0.09	0.19<0.38>0.56	0.42	0.44	0.10	0.22<0.29>0.38
45272285	0.38	0.33	0.13	0.12<0.38>0.65	0.42	0.21	0.10	0.00<0.10>0.14
34440322	0.37	0.54	0.08	0.21<0.37>0.53	0.40	0.54	0.08	0.25<0.29>0.33
45272380	0.35	0.53	0.08	0.19<0.35>0.51	0.37	0.75	0.13	0.54<0.61>0.65
10739191	0.33	0.43	0.09	0.15<0.33>0.52	0.30	0.08	0.00	0.00<0.00>0.00
23961789	0.33	0.62	0.06	0.20<0.33>0.45	0.33	-0.03	0.00	0.00<0.00>0.00
19485530	0.32	0.50	0.07	0.17<0.32>0.47	0.29	0.57	0.09	0.25<0.31>0.38
36923906	0.32	0.53	0.07	0.18<0.32>0.47	0.31	0.46	0.13	0.29<0.36>0.44
38363924	0.32	0.26	0.13	0.06<0.32>0.59	0.36	0.71	0.06	0.22<0.25>0.29
45272220	0.30	0.45	0.08	0.14<0.30>0.45	0.33	0.47	0.18	0.18<0.24>0.30
43969748	0.29	0.50	0.07	0.15<0.29>0.43	0.29	0.28	0.02	0.00>0.03>0.10
19485419	0.29	0.58	0.06	0.17<0.29>0.42	0.30	0.61	0.11	0.35<0.42>0.46
45272223	0.28	0.41	0.08	0.12<0.28>0.45	0.29	0.61	0.03	0.10<0.13>0.14
34433338	0.28	0.34	0.09	0.09<0.28>0.46	0.31	0.43	0.07	0.13<0.18>0.21
38364845	0.27	0.73	0.04	0.18<0.27>0.35	0.25	0.61	0.14	0.44<0.54>0.71

TABLE 4.2. continued

Student	<u>OLS Analysis</u>				<u>Tau Analysis</u>			
	LR slope	R2	SE slope	95% CI for R Coeff.	Theil -Sen	Tau-U	SE slope	95% CI for Theil-Sen
31333552	0.27	0.44	0.07	0.12<0.27>0.41	0.27	0.30	0.10	0.10<0.18>0.28
32837773	0.26	0.29	0.10	0.06<0.26>0.46	0.24	0.61	0.17	0.60<0.67>0.80
34433336	0.26	0.40	0.07	0.11<0.26>0.42	0.25	0.45	0.11	0.21<0.31>0.40
45272272	0.25	0.50	0.06	0.13<0.25>0.37	0.25	0.49	0.08	0.19<0.25>0.30
45272227	0.23	0.17	0.12	-0.01<0.23>.047	0.21	0.63	0.09	0.33<0.38>0.43
38363926	0.23	0.43	0.06	0.10<0.23>0.36	0.24	0.53	0.13	0.33<0.44>0.60
44299660	0.22	0.66	0.04	0.14<0.22>0.30	0.21	0.36	0.12	0.15<0.24>0.33
19485494	0.22	0.35	0.07	0.08<0.22>0.37	0.25	0.62	0.12	0.36<0.46>0.53
45272388	0.22	0.35	0.07	0.08<0.22>0.37	0.25	0.49	0.13	0.41<0.50>0.60
34440355	0.21	0.18	0.10	0.00<0.21>0.41	0.18	0.83	0.10	0.50<0.51>0.60
9840599	0.20	0.40	0.06	0.08<0.2>0.32	0.22	0.66	0.11	0.40<0.47>0.50
8597879	0.19	0.30	0.07	0.05<.019>0.33	0.20	0.65	0.10	0.40<0.43>0.50
10738098	0.19	0.20	0.09	0.01<0.19>0.38	0.22	0.70	0.21	0.80<0.94>1.00
38363813	0.18	0.40	0.05	0.07<0.18>0.28	0.18	0.46	0.09	0.20<0.27>0.33
28891518	0.18	0.13	0.11	-0.04<0.18>0.40	0.28	0.15	0.00	0.00<0.00>0.09
45272275	0.17	0.29	0.06	0.04<0.17>0.29	0.18	0.29	0.15	0.14<0.28>0.33
45272286	0.17	0.27	0.06	0.04<0.17>0.30	0.20	0.72	0.16	0.64<0.73>0.80
39024705	0.16	0.25	0.06	0.03<0.16>0.30	0.22	0.63	0.13	0.40>0.50>0.58
45272238	0.15	0.40	0.04	0.06<0.15>0.24	0.15	0.60	0.09	0.29<0.33>0.38
38363929	0.15	0.10	0.10	-0.06<0.15>0.37	0.16	0.64	0.12	0.45<0.50>0.63
45272270	0.14	0.20	0.06	0.01<0.14>0.27	0.14	0.39	0.19	0.33<0.46>0.58
18062851	0.13	0.25	0.05	0.02<0.13>0.24	0.14	0.51	0.09	0.20<0.29>0.29
18337356	0.13	0.16	0.07	-0.01<0.13>0.27	0.17	0.70	0.11	0.43<0.50>0.56
35382916	0.12	0.69	0.02	0.09<0.12>0.16	0.13	0.51	0.08	0.20<0.25>0.29

TABLE 4.2. continued

Student	<u>OLS Analysis</u>				<u>Tau Analysis</u>			
	LR slope	R2	SE slope	95% CI for R Coeff.	Theil -Sen	Tau-U	SE slope	95% CI for Theil-Sen
45272284	0.11	0.20	0.05	0.00<0.11>0.21	0.08	0.68	0.17	0.60<0.71>0.80
45272224	0.11	0.12	0.07	-0.03<0.11>0.24	0.10	0.55	0.09	0.25<0.30>0.33
45272279	0.10	0.06	0.09	-0.09<0.1>0.29	0.07	0.29	0.09	0.09<0.17>0.20
32367075	0.10	0.29	0.04	0.03<0.10>0.18	0.10	0.61	0.10	0.30<0.40>0.47
35696855	0.09	0.23	0.04	0.01<0.09>0.16	0.03	0.50	0.10	0.24<0.30>0.38
44162379	0.09	0.04	0.10	-0.12<0.09>0.30	0.07	0.37	0.06	0.09<0.14>0.18
9840591	0.08	0.06	0.07	-0.06<.08>0.22	0.00	0.32	0.11	0.14<0.22>0.30
35375408	0.04	0.04	0.05	-0.06<0.04>0.15	0.00	0.66	0.11	0.36<0.46>0.50
45272276	0.02	0.02	0.04	-0.06<0.02>0.11	0.00	0.16	0.07	0.18<0.22>0.25
38363812	0.01	0.00	0.04	-0.09<0.01>0.10	0.00	0.16	0.00	0.00<0.00>0.11
38363928	-0.01	0.01	0.04	-1.10<-0.01>0.07	0.00	0.38	0.08	0.13<0.20>0.25

Note. R2= squared Pearson's R; SESlope= standard error of the slope; results are sorted by slope size with corresponding shaded cells from large (in dark gray), to moderate (medium gray) and small (white) rates of improvement; CI= confidence interval; SE and p-values are not exact but large sample estimates only.

TABLE 4.3. Slope Differences and Significance Levels

Student	OLS Analysis			Nonparametric Analysis			Slope Diff.
	SEslope	Slope/ SEslope	Sign Level (p- value)	Theil- Sen Seslope	Slope/ SEslope	Sign Level (p- value)	If yes, which is larger? OLS or Theil-Sen
45272401	0.15	8.40	0.00	0.13	9.31	0.00	TS
45456176	0.18	6.83	0.47	0.07	17.57	0.00	TS
45454604	0.11	9.36	0.03	0.29	3.45	0.00	OLS
33067964	0.12	8.08	0.01	0.50	8.80	0.00	TS
45272399	0.16	5.56	0.96	0.25	3.00	0.00	OLS
45272229	0.09	9.67	0.31	0.20	4.25	0.00	TS
18822785	0.13	6.00	1.00	0.07	10.14	0.00	TS
45272400	0.13	6.00	0.28	0.19	3.95	0.00	OLS
33662217	0.13	5.46	0.01	0.16	4.19	0.00	OLS
31982440	0.08	8.63	0.29	0.17	4.29	0.00	OLS
45272390	0.16	3.81	0.01	0.13	2.92	0.00	OLS
31333553	0.07	8.57	0.28	0.11	4.64	0.00	OLS
38929384	0.06	10.00	0.79	0.07	8.71	0.01	OLS
46195407	0.15	3.80	0.08	0.12	2.75	0.00	OLS
45272278	0.12	4.75	0.00	0.13	4.54	0.00	OLS
46542501	0.09	6.11	0.58	0.13	3.85	0.00	OLS
45272391	0.07	7.86	0.00	0.10	5.50	0.01	OLS
45272382	0.09	6.00	0.01	0.13	3.62	0.00	OLS
34433393	0.11	4.91	0.52	0.09	6.00	0.00	TS
34429246	0.11	4.82	0.04	0.11	4.55	0.00	OLS

TABLE 4.3. continued

Student	OLS Analysis			Nonparametric Analysis			Slope Diff.
	SEslope	Slope/ SEslope	Sign Level (p- value)	Theil- Sen Seslope	Slope/ SEslope	Sign Level (p- value)	If yes, which is larger? OLS or Theil-Sen
45724743	0.08	6.63	0.61	0.10	5.70	0.00	OLS
45272398	0.17	3.00	0.04	0.11	2.27	0.00	OLS
45272381	0.09	5.56	0.00	0.08	6.25	0.01	TS
33058186	0.08	6.13	0.00	0.05	10.00	0.10	TS
45272396	0.13	3.77	0.00	0.14	3.57	0.00	OLS
45272394	0.18	2.67	0.01	0.04	5.00	0.02	TS
33067965	0.06	8.00	0.00	0.00	0.00	0.47	OLS
33730300	0.08	5.75	0.01	0.07	6.57	0.01	TS
27523261	0.06	7.67	0.00	0.17	2.94	0.00	OLS
38363965	0.07	6.57	0.00	0.07	7.14	0.30	TS
45272385	0.10	4.50	0.02	0.10	3.3	0.10	OLS
9840636	0.07	6.43	0.02	0.08	5.75	0.00	OLS
19485963	0.13	3.38	0.00	0.07	6.57	0.06	TS
32921873	0.09	4.78	0.00	0.08	4.75	0.01	OLS
43889271	0.09	4.67	0.17	0.10	4.30	0.00	OLS
33067969	0.10	4.20	0.02	0.10	4.40	0.00	TS
44922885	0.06	6.83	0.92	0.05	8.00	0.00	TS
43957949	0.03	13.67	0.00	0.10	4.10	0.00	OLS
32085739	0.07	5.86	0.01	0.10	4.30	0.03	OLS
19311652	0.07	5.71	0.01	0.08	5.00	0.40	OLS
45272237	0.07	5.57	0.00	0.08	4.88	0.00	OLS

TABLE 4.3. continued

Student	OLS Analysis			Nonparametric Analysis			Slope Diff.
	SEslope	Slope/ SEslope	Sign Level (p- value)	Theil- Sen Seslope	Slope/ SEslope	Sign Level (p- value)	If yes, which is larger? OLS or Theil-Sen
45272277	0.09	4.33	0.07	0.10	4.10	0.00	OLS
36924182	0.09	4.22	0.01	0.10	4.20	0.00	OLS
45272285	0.13	2.92	0.00	0.10	4.20	0.18	TS
34440322	0.08	4.63	0.00	0.08	5.00	0.00	TS
45272380	0.08	4.38	0.00	0.13	2.85	0.00	OLS
10739191	0.09	3.67	0.00	0.00	0.00	0.63	OLS
23961789	0.06	5.50	0.00	0.00	0.00	0.86	OLS
19485530	0.07	4.57	0.00	0.09	3.22	0.00	OLS
36923906	0.07	4.57	0.05	0.13	2.38	0.00	OLS
38363924	0.13	2.46	0.04	0.06	6.00	0.00	TS
45272220	0.08	3.75	0.00	0.18	1.83	0.00	OLS
43969748	0.07	4.14	0.00	0.02	14.50	0.08	TS
19485419	0.06	4.83	0.02	0.11	2.73	0.00	OLS
45272223	0.08	3.50	0.00	0.03	9.67	0.00	TS
34433338	0.09	3.11	0.00	0.07	4.43	0.01	TS
38364845	0.04	6.75	0.09	0.14	1.79	0.00	OLS
31333552	0.07	3.86	0.00	0.10	2.70	0.06	OLS
32837773	0.10	2.60	0.00	0.17	1.41	0.00	OLS
34433336	0.07	3.71	0.00	0.11	2.27	0.00	OLS
45272272	0.06	4.17	0.01	0.08	3.13	0.00	OLS
45272227	0.12	1.92	0.00	0.09	2.33	0.00	TS

TABLE 4.3. continued

Student	OLS Analysis			Nonparametric Analysis			Slope Diff.
	SEslope	Slope/ SEslope	Sign Level (p- value)	Theil- Sen Seslope	Slope/ SEslope	Sign Level (p- value)	If yes, which is larger? OLS or Theil-Sen
38363926	0.06	3.83	0.02	0.13	1.85	0.00	OLS
44299660	0.04	5.50	0.00	0.12	1.75	0.02	OLS
19485494	0.07	3.14	0.02	0.12	2.08	0.00	OLS
45272388	0.07	3.14	0.00	0.13	1.92	0.00	OLS
34440355	0.10	2.10	0.01	0.10	1.80	0.00	OLS
9840599	0.06	3.33	0.00	0.11	2.00	0.00	OLS
8597879	0.07	2.71	0.04	0.10	2.00	0.00	OLS
10738098	0.09	2.11	0.15	0.21	1.048	0.00	OLS
38363813	0.05	3.60	0.01	0.09	2.00	0.00	OLS
28891518	0.11	1.64	0.00	0.00	0.00	0.35	OLS
45272275	0.06	2.83	0.00	0.15	1.20	0.07	OLS
45272286	0.06	2.83	0.10	0.16	1.25	0.00	OLS
39024705	0.06	2.67	0.27	0.13	1.69	0.00	OLS
45272238	0.04	3.75	0.00	0.09	1.67	0.00	OLS
38363929	0.10	1.50	0.05	0.12	1.33	0.00	OLS
45272270	0.06	2.33	0.00	0.19	0.74	0.01	OLS
18062851	0.05	2.60	0.82	0.09	1.56	0.00	OLS
18337356	0.07	1.86	0.00	0.11	1.55	0.00	OLS
35382916	0.02	6.00	0.00	0.08	1.63	0.00	OLS
45272284	0.05	2.20	0.34	0.17	0.47	0.00	OLS
45272224	0.07	1.57	0.01	0.09	1.11	0.00	OLS

TABLE 4.3. continued

Student	OLS Analysis			Nonparametric Analysis			Slope Diff.
	SEslope	Slope/ SEslope	Sign Level (p- value)	Theil- Sen Seslope	Slope/ SEslope	Sign Level (p- value)	If yes, which is larger? OLS or Theil-Sen
45272279	0.09	1.11	0.00	0.09	0.78	0.07	OLS
32367075	0.04	2.50	0.02	0.10	1.00	0.00	OLS
35696855	0.04	2.25	0.66	0.10	0.30	0.00	OLS
44162379	0.10	0.90	0.09	0.06	1.17	0.02	TS
9840591	0.07	1.14	0.00	0.11	0.00	0.04	OLS
35375408	0.05	0.80	0.11	0.11	0.00	0.00	OLS
45272276	0.04	0.50	0.00	0.07	0.00	0.00	OLS
38363812	0.04	0.25	0.00	0.00	0.00	0.11	OLS
38363928	0.04	-0.25	0.00	0.08	0.00	0.02	TS

Note. SEslope= standard error of the slope; SE and p-values are not exact but large sample estimates only; Slope Diff.= comparison of slope/SEslope ratio indicating if there was a difference in the ratio and if yes, which was larger.

TABLE 4.4. Tests of Residuals

Parametric Analysis									
Student	Shapiro- Wilk	Modif. Levene	r_{auto} (lag -1)	Durbin Watson	Student	Shapiro- Wilk	Modif. Levene	r_{auto} (lag -1)	Durbin Watson
45724743	0.97	0.22	-0.23	2.11	38363812	0.78	3.51	-0.37	2.70
46195407	0.69	1.58	-0.14	2.19	38363928	0.77	0.21	0.07	1.75
45456176	0.95	0.38	-0.36	2.70	36923906	0.95	0.79	-0.16	2.26
45272238	0.85	0.01	-0.09	2.15	38363924	0.77	3.48	-0.22	2.43
45272401	0.95	5.62	0.30	1.12	38364845	0.93	0.17	-0.06	2.09
45454604	0.88	0.42	-0.25	2.47	38363926	0.96	0.06	-0.15	2.26
45272398	0.66	0.95	-0.03	1.29	35696855	0.97	1.29	0.42	0.77
45272400	0.89	2.23	0.11	1.63	36924182	0.82	0.57	0.05	1.83
45272396	0.96	2.09	0.35	1.30	35382916	0.96	0.55	0.24	1.48
45272399	0.80	1.60	-0.29	2.53	38363813	0.90	0.02	0.20	1.59
43889271	0.98	0.05	0.03	1.89	34433393	0.99	2.21	-0.28	2.50
45272382	0.92	0.64	0.16	1.24	34440355	0.94	0.02	0.40	1.03
45272394	0.73	0.90	0.09	1.12	33662217	0.94	0.50	-0.25	2.43
38363965	0.97	0.63	-0.20	2.31	34433338	0.96	4.03	-0.07	2.02
45272391	0.95	1.18	0.28	1.42	34433336	0.92	0.56	-0.05	1.98
45272381	0.97	5.80	-0.04	2.08	32921873	0.88	1.67	0.12	1.32
45272388	0.95	0.00	0.14	1.60	33067969	0.97	0.79	0.31	1.27
45272277	0.98	1.26	-0.65	3.23	32837773	0.96	0.62	-0.05	2.07
45272385	0.87	3.58	0.31	0.85	33730300	0.98	0.92	-0.07	2.01
45272380	0.98	0.13	-0.14	2.26	34429246	0.93	0.28	-0.21	2.32
45272390	0.72	1.31	-0.13	1.53	31333553	0.92	4.27	0.16	1.66
34440322	0.93	6.77	-26.00	2.45	33067965	0.97	0.52	0.47	0.92
45272286	0.91	3.00	-0.37	2.73	32085739	0.92	3.88	0.21	1.19

TABLE 4.4. continued

Parametric Analysis									
Student	Shapiro- Wilk	Modif. Levene	r _{auto} (lag -1)	Durbin Watson	Student	Shapiro- Wilk	Modif. Levene	r _{auto} (lag -1)	Durbin Watson
45272284	0.97	0.30	-0.23	2.29	33067964	0.96	0.44	0.20	1.54
45272285	0.87	2.60	-0.05	2.04	31333552	0.94	0.51	0.16	1.55
32367075	0.95	0.06	-0.01	2.00	35375408	0.97	0.16	-0.46	2.91
45272276	0.90	0.66	-0.15	2.29	28891518	0.96	2.63	0.25	1.42
45272275	0.93	0.38	-0.13	2.20	31982440	0.94	0.25	-0.10	2.18
45272278	0.96	0.03	0.27	1.41	33058186	0.98	2.98	-0.37	2.55
45272279	0.91	0.39	0.11	1.64	23961789	0.91	0.46	-0.29	2.35
38363929	0.89	0.25	0.44	1.10	26542501	0.96	0.34	0.10	1.46
45272272	0.97	1.05	-0.43	2.83	19485963	0.96	0.69	0.48	0.96
45272270	0.96	2.44	-0.26	2.47	19485530	0.96	1.67	0.08	1.60
39024705	0.94	3.39	0.10	1.79	27523261	0.96	5.09	-0.16	2.09
45272229	0.97	6.27	-0.08	2.07	19485494	0.88	0.29	0.26	1.38
45272237	0.94	0.26	0.06	1.79	18822785	0.91	9.59	-0.51	3.02
44299660	0.97	1.05	0.13	1.74	19485419	0.95	1.05	-0.33	2.59
45272220	0.94	0.15	-0.42	2.78	18337356	0.93	1.12	-0.35	2.62
45272227	0.90	1.27	0.19	1.54	19311652	0.93	0.01	-0.27	2.49
44162379	0.93	0.05	0.53	0.86	10739191	0.92	0.46	-0.25	2.50
43957949	0.98	0.15	0.10	1.70	18062851	0.92	0.54	-0.16	2.28
44922885	0.97	0.20	-0.09	2.13	10738098	0.96	0.62	0.25	1.21
45272223	0.97	0.06	0.16	1.62	9840636	0.98	5.13	-0.03	2.05
45272224	0.95	3.80	0.25	1.45	9840599	0.85	0.31	-0.40	2.76
43969748	0.91	0.03	-0.13	2.02	9840591	0.96	0.24	-0.10	2.17
38929384	0.92	0.48	-0.10	2.16	8597879	0.98	3.03	-0.30	2.59

Note. SEslope= standard error of the slope; r_{auto}= autocorrelation or serial dependence; Durbin Watson- test of autocorrelation; Meet Parametric Assump.= meets assumptions of normality and equal variance.

VITA

Susan Adele Dupise Bruhl

Address: Addison Northeast Supervisory Union, 72 Munsil Avenue, Bristol, VT 05443
Email: sbruhl@anesu.org

Education

Master of Education- University of Hawaii at Manoa, 1994
Bachelor of Fine Arts- James Madison University, 1989

Professional Experience

Special Education Director

Addison Northeast Supervisory Union, Bristol, VT 2006- present

Assistant Professor and Field Supervisor

College of Saint Joseph, Education Division, Rutland, VT, 2003-2005
Texas A&M University, Department of Educational Psychology, College Station TX, 1999-2003.

Classroom Teacher

St. Mary's School, Middlebury, VT, 2005-2006
College Station ISD, College Station, TX, 1994-1999
Kailua Intermediate, Kailua, HI, August- December 1991.

Publications

Lynch, P. & Bruhl, S. (2005). Inclusive practices in secondary settings. In P. Zionts (Ed.) *Inclusion Strategies for Students with Learning and Behavioral Problems*, (2nd ed.) (pp. 93-109). Austin, TX: PRO-ED, Inc.

Bruhl, S.D., Callicott, K.J. & Fournier, C.J. (2003). *Real Kids, Real Teaching: A Handbook for the Journey* (2nd ed.). Boston: Pearson.

Prater, M. A., Bruhl, S., & Serna, L.A. (1998). Acquiring Social Skills Through Cooperative Learning and Directed Teacher Instruction. *Remedial & Special Education*, 19 (3), 160-172.

Honors and Awards

Recipient of the Regent's Scholarship for Doctoral Studies 2003-2004
 Teacher of the Year, College Station Junior High 1998