An Evaluation of the Accuracy of Time Series Interpretations of CBM-R Progress Monitoring Data

A DISSERTATION
SUBMITTED TO THE FACULTY OF
THE UNIVERSITY OF MINNESTOA
BY

Ethan Richard Van Norman

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Theodore J. Christ, Advisor

June, 2015

Acknowledgements

I wrote this section last. I spent too much time trying to poetically convey how thankful I am for all of the support I have received throughout my life. Rather than share an anecdote I opted to list the few among many who helped me achieve what I have thus far and those that inspire me to achieve more.

The first person is my mother Jeanine Wilker. Dr. Wilker is not only an exemplar educator she is an exemplar human being. She raised two sons while working full time and earning two graduate degrees. I hope that I will not only replicate her commitment and passion for education but also her unwavering dedication to her family.

The second person is my undergraduate advisor Dr. Glenn Reeder. Dr. Reeder is in large part the reason I decided to pursue a graduate education. He provided the initial mentorship that led me to believe that research was not all that bad.

The third person is my graduate advisor Dr. Theodore Christ. I believe I have achieved what I have thus far is in large part because of the high expectations placed upon me by Dr. Christ. In the midst of my graduate career I would often become disgruntled with the near perfectionist standards he held me to. In time I realized that these standards are what molded me into the scientist I am today. I am forever indebted to Dr. Christ for the experiences he afforded me, the responsibilities he entrusted me with, and the respect he showed me when he identified me as a colleague.

Last I must acknowledge the friends I made during my graduate career. Without them I cannot imagine how much more difficult the last five years would have been. I value the life long connections I made with the individuals I met at the University of Minnesota above anything else.

Abstract

Curriculum based measurement of reading (CBM-R) is used to monitor the effects of academic interventions for individual students. Decisions to continue, modify, or terminate instructional programs are made by interpreting patterns of observations collected across time. Educators visually analyze or apply decision rules to evaluate student progress. Despite the popularity of CBM-R as a progress monitoring tool, there is a paucity of research evaluating the accuracy of visual analysis and decision rules. Inaccurate interpretations undermine the use of CBM-R as a progress monitoring tool because educators may continue ineffective interventions or prematurely terminate effective interventions. The accuracy of visual analysis and decision rules were investigated in this project. In Study 1 a large extant dataset was analyzed to identify measurement characteristics of CBM-R progress monitoring data. In Study 2 the accuracy of visual analysis and decision rules were evaluated by comparing responses from visual analysts and decision rules with responses of an expert panel. One hundred eight progress monitoring graphs were evaluated in Study 2. The manner in which progress monitoring graphs differed was informed by the results of Study 1. The results of this project suggest evaluation method, number of weeks data are collected, variability of observations, and whether the student is making adequate progress influence the probability of correct decisions. Educators and researchers can improve the probability of correct decisions by visually analyzing progress monitoring graphs with a goal line and trend line, minimizing variability, and collecting data for longer than six weeks. The implications of the findings, limitations, and needs for future research are discussed.

Table of Contents

Acknowledgements ... i
Abstract ... ii
List of Tables ... vi
List of Figures ... vii

Chapter 1 - Introduction ... 1
CBM-R Progress Monitoring ... 1
CBM-R Decision Rules ... 3
CBM-R Simulation Studies ... 4
Unresolved Issues With CBM-R Decision Rules ... 5
Purpose ... 6
Definitions Of Key Terms ... 6

Chapter 2 – Review Of The Literature ... 8
Interpretative Argument ... 10
Recent CBM-R IUA Efforts ... 10
Literature Review ... 15
Inclusion Criteria ... 16
Search Strategies ... 16
Decision Rule Studies ... 18
Visual Analysis Study ... 22
Summary ... 23
Conclusion ... 24

Chapter 3 – Study 1 ... 25
CBM-R as a Progress Monitoring Tool ... 25
Characteristics of Progress Monitoring Data ... 27
Duration ... 27
Missing Data ... 29
Variability ... 30
Aberrant Values ... 31
Growth Patterns ... 32
Intercept and Slope ... 33
Autocorrelation ... 34
Purpose ... 35
Research Questions ... 35
Methods ... 35
Participants ... 35
Procedure ... 36
Analyses ... 36
Duration ... 39
Missing Data ... 39
Variability ... 39
Aberrant Values ... 40

Growth Patterns 40
Intercept And Slope 40
Autocorrelation 41
Results...41
Duration 41
Missing Data 44
Variability 46
Aberrant Values 47
Growth Patterns 51
Intercept And Slope 57
Autocorrelation 61
Discussion 63
Characteristics of Progress Monitoring Data 63
Duration 63
Missing Data 64
Variability 64
Aberrant Values 65
Growth Patterns 65
Intercept And Slope 66
Autocorrelation 67
Validity Of Simulation Studies 67
Implications For Future Research 70
Limitations 71
Conclusions...72

Chapter 4 – Study 2..75
Curriculum Based Measurement...75
Visual Analysis 76
Visual analysis and cbm-r 78
Decision Rules 79
Purpose..82
Research questions..82
Methods...83
Participants...83
Expert panel 83
Visual analysts 84
Materials 84
Design 85
Evaluation method 86
Trend magnitude 86
Duration 86
Variability 87
Correct responses 87
Procedure 88
Expert panel 88

Visual analysts 89
Decision rules 91
Analyses 91
Descriptive 92
Inferential 92
Results 93
Visual analysts 95
Decision rules 100
Comparison of methods 103
Discussion 105
Trend magnitude 106
Duration 106
Variability 107
Implications For Practice 107
Implications For Research 109
Future research 110
Limitations 111
Conclusion 113

Chapter 5 – Integrated Discussion 116
Study 1 117
Study 2 119
Conclusion 121

References 122
Appendix A Descriptive Analyses 136
Appendix B Original Inferential Analyses 143

List of Tables

Table 1: Parameter Estimates for Progress Monitoring Data from Christ Zopluoglu, Monaghen, & Van Norman, 2013..38

Table 2: Descriptive Statistics of Characteristics of Progress Monitoring Cases (n=3,078) 42

Table 3: Summary of the Number and Proportion of Progress Monitoring Cases with Aberrant Values 53

Table 4: Frequency and Proportion of Cases with Aberrant Values Conditioned on Duration 54

Table 5: Frequency and Proportion of Cases with and without Aberrant Values Conditioned on the Magnitude of Variability of Observations 55

Table 6: Multilevel Analysis Predicting Log Odds of a Correct Decision for Visual Analysts..98

Table 7: Logistic Regression Analysis Predicting Log Odds of a Correct Response for Decision Rules 102

Table 8: Comparison of the Average Probability of a Correct Decision across Evaluation Methods and Progress Monitoring Conditions 104

Table 9: Agreement Indices for Visual Analysts 140

Table 10: Agreement Indices for Decision Rules 147

Table 11: Original Multilevel Analysis Predicting Log Odds of a Correct Decision for Visual Analysts 149

Table 12: Original Logistic Regression Analysis Predicting Log Odds of a Correct Response for Decision Rules 150

Table 13: Average Probabilities of a Correct Decision for each Evaluation Method 151

List of Figures

Figure 1: An example of continuously evaluating changes in a student's reading instruction. Reprinted from Deno (1990)...2

Figure 2: Example interpretative use argument for using oral reading rate as an indicator of broad reading competence. From Christ, Nelson, and Van Norman (2014). 12

Figure 3: Number of cases for each progress monitoring duration schedule measured in weeks (n = 3,078)..43

Figure 4: Box and whisker plots summarizing the percentage of missing observations within progress monitoring cases (n=3,078) conditioned on the duration of data collection schedules 45

Figure 5: Box and whiskers plot summarizing standard error of the estimate values for progress monitoring cases (n = 3,078) conditioned on the duration of data collection schedules 49

Figure 6: Box and whiskers plot summarizing standard error of the slope values for progress monitoring cases (n = 3,078) conditioned on the duration of data collection schedules 50

Figure 7: Box and whisker plots summarizing unstandardized week * week regression terms for progress monitoring cases conditioned on the duration of data collection 56

Figure 8: Box and whisker plots summarizing intercept estimates (predicted words read correct per minute score at the beginning of data collection) for progress monitoring cases (n=3,078) conditioned on the duration of data collection schedules........................59

Figure 9: Box and whisker plots summarizing slope estimates (average weekly increase in words read correct per minute) of progress monitoring cases (n = 3,078) conditioned on the duration of progress monitoring schedules. 60

Figure 10: Box and whisker plots summarizing the degree of lag-1 autocorrelation in progress monitoring cases (n = 3,078) conditioned on the duration of data collection schedules. 62

Chapter 1

Introduction

Effective educators continuously evaluate the impact of interventions they deliver to students (Fuchs & Fuchs, 1986). Progress monitoring is the act of collecting assessment data continuously across time to formatively assess instructional programming (see the end of this chapter for definitions of key terms; Deno, 1990). Educational professionals use progress monitoring data to inform decisions to continue, modify, or terminate academic and behavioral interventions for individual students. Researchers frequently state that academic achievement improves if instruction is changed formatively when a student is showing a lack of progress (Fuchs & Fuchs, 1986; Stecker, Fuchs, & Fuchs, 2005). Yet, in 2009 a task force of educational experts commissioned by the federal government found mixed results for the effects of progress monitoring. Stated differently, the panel concluded that the empirical evidence for progress monitoring practices were weak (Gersten et al., 2009). Despite the lack of empirical evidence for progress monitoring, the panel supported the act of measuring instructional effects as put forward by Deno (1990) and Fuchs and Fuchs (1986). One of the most commonly used procedures to monitor student progress is curriculum based measurement (CBM; Deno, 1985). Further, CBM of oral reading (CBM-R) is the most commonly used and widely researched progress monitoring procedure (Reschly, Busch, Betts, Deno, & Long, 2009; Wayman, Wallace, Wiley, Ticha, & Espin, 2007).

CBM-R Progress Monitoring

To monitor student progress with CBM-R, students read grade level passages aloud for one minute while the administrator calculates the number of words read correctly per minute (WRCM). Educators administer alternate forms across time and plot the observations on a time series graph (see Figure 1). Educators then evaluate the pattern of observations to make decisions about a student’s instructional programming (Deno, 1986). The student’s rate of improvement (ROI), or growth in oral reading rate, is used to evaluate instructional effects. A tangible goal is a useful standard to which the student’s ROI is compared (Shapiro, 2008; vanDerHeyden, Witt, & Barnett, 2005). An expected ROI, or goal line, defines the magnitude of expected progress (Shapiro, 2011).

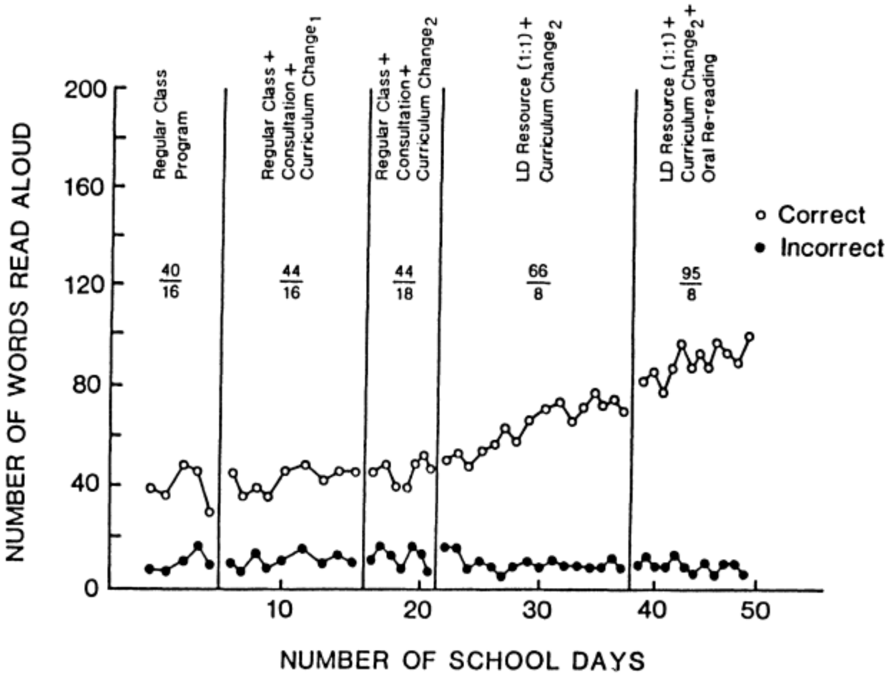

Figure 1. An example of continuously evaluating changes in a student's reading instruction. Reprinted from Deno (1990).

CBM-R Decision Rules

Educators often make instructional decisions based on a student's performance in relation to a goal line. If the computed slope, or trend line, of CBM-R observations is less steep than the goal line after a pre-determined number of weeks, the teacher is prompted to change the student's instruction (Shinn, 1989). Alternatively if the three, four, or five most recent observations fall below the goal line, the teacher is prompted to change the student's instruction (White & Haring, 1980). These methods are referred to as trend line and data point decision rules respectively (Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013). When educators evaluate CBM-R progress monitoring data, they must not only evaluate *if* an intervention is effective, but they must determine *how* effective an intervention is. If a student is making progress but the ROI is so low then that the student is unlikely to reach their end of year goal, an instructional modification should be made. Using decision rules, in conjunction with a goal line, theoretically, adds some level of structure and consistency to decisions to modify interventions (Shinn, 2008).

Ample evidence suggests that oral reading rate is a robust indicator of broad reading competence (Reschly, et al., 2009; Shinn, Good, Knutson, Tilly, & Collins, 1992; Wayman et al., 2007). However, growth estimates based upon CBM-R progress monitoring data are often unreliable indices of instructional effects over short periods of time (Christ, 2006; Good & Shinn, 1990). Early research suggested that CBM-R was sensitive to short term improvement (Deno, Marston, & Tindal, 1986). Yet, more recent evidence suggests that CBM-R data collected in less than ideal situations with sub-par instruments yield highly unstable short term estimates of growth (Christ, 2006; Hintze & Christ, 2004). To complicate matters, in a recent review of the literature on CBM-R decision rules, not a single study identified in the review evaluated the accuracy of decision based on trend line or data point decision rules (Ardoin et al., 2013). The lack of research is a major cause for concern considering the potential resources necessary to collect CBM-R progress monitoring data. In addition, progress monitoring is used to refine and optimize instruction. If visual analysis and CBM-R decision rules yield inaccurate decisions then it might confer harm to students.

CBM-R Simulation Studies

Within their review, Ardoin and colleagues (2013) stated that recent simulation studies offer the closest evaluation of the accuracy of CBM-R decision rules. However, the studies were too recent to be included in the review, as where several other unpublished manuscripts. Those studies were simulations in which, a large extant dataset was analyzed to derive model parameters. After that, true growth was specified, and observed growth was generated across different progress monitoring scenarios (e.g., Christ, Zopluoglu, Long, & Monaghen, 2012; Christ, Zopluoglu, Monaghen, & Van Norman, 2013). Those scenarios included: the quality of instrumentation and degree of standardization within data collection conditions (the residual variance or variability of observations), the number of weeks progress was monitored, the number of times per week data were collected, the number of observations collected during each administration, and the manner in which observed growth was summarized and decision rules were applied.

A more comprehensive review of CBM-R progress monitoring simulation studies is offered in Chapter 2. In brief, after generating true growth and observed growth across hundreds of conditions, the researchers assessed the correspondence between true growth and observed growth, the precision of observed growth estimates, and the reliability of observed growth estimates (Christ, Zopluoglu, Long et al., 2012). In other studies researchers evaluated the ideal magnitude of observed growth to predict whether or not a student was improving adequately in relation to a goal line (Christ, Zopluoglu, Monaghen et al., 2013). Despite the benefit of simulation studies, particularly in the ability to evaluate hundreds of progress monitoring scenarios without expending the time, money, and resources to collect that data, several issues remain unresolved.

Unresolved Issues with CBM-R Decision Rules

The authors of the simulation studies analyzed the same extant dataset to estimate model parameters for simulations in each study. It is unclear if model parameters, and potentially the findings of each study, would differ if another large extant dataset were analyzed. Moreover, the authors did not present an in-depth exploratory analysis of the extant dataset from which model parameters were derived. As with any simulation study, the validity of the results depend on the degree to which underlying assumptions associated with the model and model parameters were met.

In each of the simulation studies the authors highlight the fact that CBM-R progress monitoring data are meant for use within an idiographic framework. That is educators evaluate growth and make decisions at the individual student level. Yet the analyses presented in the studies were based on thousands of simulated cases. That is, reliability estimates, diagnostic accuracy statistics, and the level of precision between observed and true growth for each progress monitoring scenario were presented as averages based upon thousands of simulated cases. The degree to which the accuracy of decisions converges with instances where a single student is evaluated without a psychometrically strong criterion for true growth is unclear. Further, the simulations represent a situation in which treatment decisions are made automatically. Practitioners that use single case design methodology rarely use statistical methods in isolation to interpret progress monitoring data. Instead, educators act as visual analysts and evaluate the level, trend, and variability of observations with statistical and graphic aids to ultimately make treatment decisions. Researchers have yet to fully evaluate decision accuracy as it relates to visual analysis and CBM-R decision rules. There are many questions left unanswered.

It stands to reason that the literature on visual analysis and decision rules within the context of CBM-R could benefit from studies that: (a) explore characteristics of a large extant dataset to identify patterns of measurement characteristics, (b) evaluate the accuracy of visual analysis and decision rules using extant (non-simulated) student data, and (c) compare the accuracy of visual analysts and decision rules.

Purpose

This project consisted of two related studies. The purpose of Study 1 was to explore the measurement characteristics of a large extant CBM-R progress monitoring dataset. The purpose of Study 2 was to evaluate the accuracy of visual analysis and CBM-R decision rules across conditions identified in Study 1. More specific research questions for Study 1 and Study 2 can be found in Chapters 3 and 4 respectively.

Definitions of Key Terms

The following definitions of technical terms are provided to aid the reader:

Accuracy: The degree to which outcomes from visual analysis and CBM-R decision rules agree with interpretations from an expert panel.

Autocorrelation: The degree to which one observation is related to (or predictive of) future observations in a time series.

Curriculum based measurement: A class of assessments created by Deno and colleagues to index a student's level and rate of improvement, or growth, in general academic skill areas.

Data point decision rule: A method of interpreting CBM progress monitoring data where the last 3, 4, or 5 observations are evaluated in relation to an expected rate of improvement or goal line to guide treatment decisions.

Graphic aids: Tools that summarize patterns on time series graphs to assist practitioners interpret intervention effects.

Progress monitoring: The act of continuously measuring instructional effects across time.

Standard error of the estimate: The average level of spread of observations around a line of best fit. A measure of precision of point estimates in a time series.

Standard error of the slope: A measure of precision of growth estimates from time series data.

Trend line decision rule: A method of interpreting CBM progress monitoring data where the slope of a line fitted through collected observations is evaluated in relation to a goal line to guide treatment decisions.

Visual analysis: The process of interpreting level, trend, and variability of time series data for a single subject to evaluate the effects of an intervention.

Chapter 2

Review of the Literature

Deno (1985; 2003) and colleagues developed CBM to measure the effects of special education instruction for individual students across relatively brief periods of time. To do so adequately they argued a tool had to be defensible, flexible, repeatable, and efficient. Broadly speaking, an assessment is considered defensible when it possesses strong psychometric evidence for its intended use (measuring instructional effects across time) and evidence that using the tool results in improved outcomes for relevant stakeholders (academic achievement improves by matching students to appropriate interventions; Messick, 1995). A series of quasi-experimental studies conducted in the late 1980's and early 1990's suggested that when educators use CBM-R in conjunction with decision rules to formatively assess instructional programs, students, on average, experience greater academic gains than those who did not receive similar data based programing (e.g., Fuchs, Fuchs, & Hamlett, 1989; Fuchs, Fuchs, Hamlett, & Ferguson, 1992). This line of inquiry is often used to tangentially affirm the accuracy of existing CBM-R decision rules and visual analytic methods. A rival hypothesis is that educators who use CBM-R and actively interpret time series data are more likely to engage in certain behaviors (e.g., using more specific goals, using other more objective measures of student achievement, providing feed back to students, using performance incentives, changing instruction more frequently) than teachers who are not using similar data based programing (Ardoin et al., 2013). In turn, those behaviors may be affecting superior achievement, regardless of whether interpretations of progress monitoring data are accurate.

A more contemporary psychometric framework may provide a more compelling context to understand the implications of not directly investigating CBM-R decision rules and visual analytic methods. Kane (2006; 2012) proposed an argument-based approach to validation. The framework is an outgrowth of the work of Cronbach (1947) and Messick (1989). Within the argument based-approach to validity Kane reminds test developers and users that tests themselves are not valid or invalid, proposed uses are. Consequently, the validity of a proposed interpretation and use of a test is determined by the plausibility of the claims being made.

Kane (2013) suggested that test users and developers generate an interpretation/use argument (IUA) to organize the network of inferences and assumptions that must take place before a decision based on a test performance can be made. After the IUA is laid out, a second argument is generated, the validity argument. Within the validity argument, an individual evaluates the plausibility of the IUA. An interpretation or use is thus considered valid if the IUA is coherent and complete and the assumptions are plausible a-priori– or are supported with empirical evidence.

The purpose of this chapter was to explain how the current project contributes to the validity argument when using CBM-R to measure instructional effects. First, the specific interpretative argument assumed for this project will be presented. Next, recent efforts to outline the inferences and assumptions associated with using CBM-R as progress monitoring tool will be discussed. After that, existing studies that address the plausibility of the interpretative argument will be reviewed. The final section of the chapter will address how Study 1 and Study 2 contribute to assessing the plausibility of the initial interpretative argument.

Interpretative Argument

When considering CBM-R as a progress-monitoring tool, a specific interpretative argument is that changes in oral reading scores across time are sufficiently representative of and sensitive to improvement in reading achievement. One important assumption then is that evaluations from decision rules and interpretations using visual analysis are in fact accurate. Stated differently, when CBM-R data are plotted graphically, decisions regarding student progress reflect actual changes in student performance. Evaluating the accuracy of time-series interpretations of CBM-R progress monitoring data is a direct precursor to consequential validity. Decisions based on the trajectory of CBM-R time series data have positive and negative consequences for students, educators, and school systems. Incorrect decisions undermine the validity of using CBM-R to measure instructional effects. Conversely, correct decisions strengthen the argument for the validity and use of CBM-R. Before the validity of the argument - methods to interpret CBM-R time series data are in fact accurate, can be assessed -the inferences and assumptions associated with using CBM-R to measure instructional effects are reviewed.

Recent CBM-R IUA Efforts

Christ, Van Norman, and Nelson (2014) recently outlined numerous inferences and assumptions associated with using CBM-R as a measure of general reading achievement consistent with the argument based approach to validation (see Figure 2). The assumptions and inferences they proposed are reviewed briefly here.

In the upper left corner of the figure, Christ and colleagues begin with a trait. Broadly speaking, the term trait refers to a latent ability that influences or causes behavior in a target domain (the cloud in the figure). Ability in this sense, as the term latent trait implies, is not directly observable. As a result, to estimate the level or state of literacy a set of behaviors that are observable to infer the state must be defined. A fundamental assumption is that behavior in the target domain is influenced by the level of the trait, and not overly influenced by irrelevant sources of variance. The authors chose a cloud shape to depict the target domain to convey that defining a target domain is extremely challenging and target domains are very difficult to fully observe. If the target domain is conceptualized as broad reading achievement the number of possible behaviors that could be indicative of or influenced by literacy is staggering. As a result, target domains rarely consist of all possible behaviors in all possible contexts that are reflective of the trait of interest (Brennan, 2000). Instead the target domain is restricted to a testable domain (indicated by the solid box within the cloud). This testable domain can be referred to as a universe of generalization. From the universe of generalization tasks and contexts are sampled from an assessment. When using CBM-R, the universe of generalization consists of orally reading narrative grade level passages.

Hypothesized Empirical Relationships

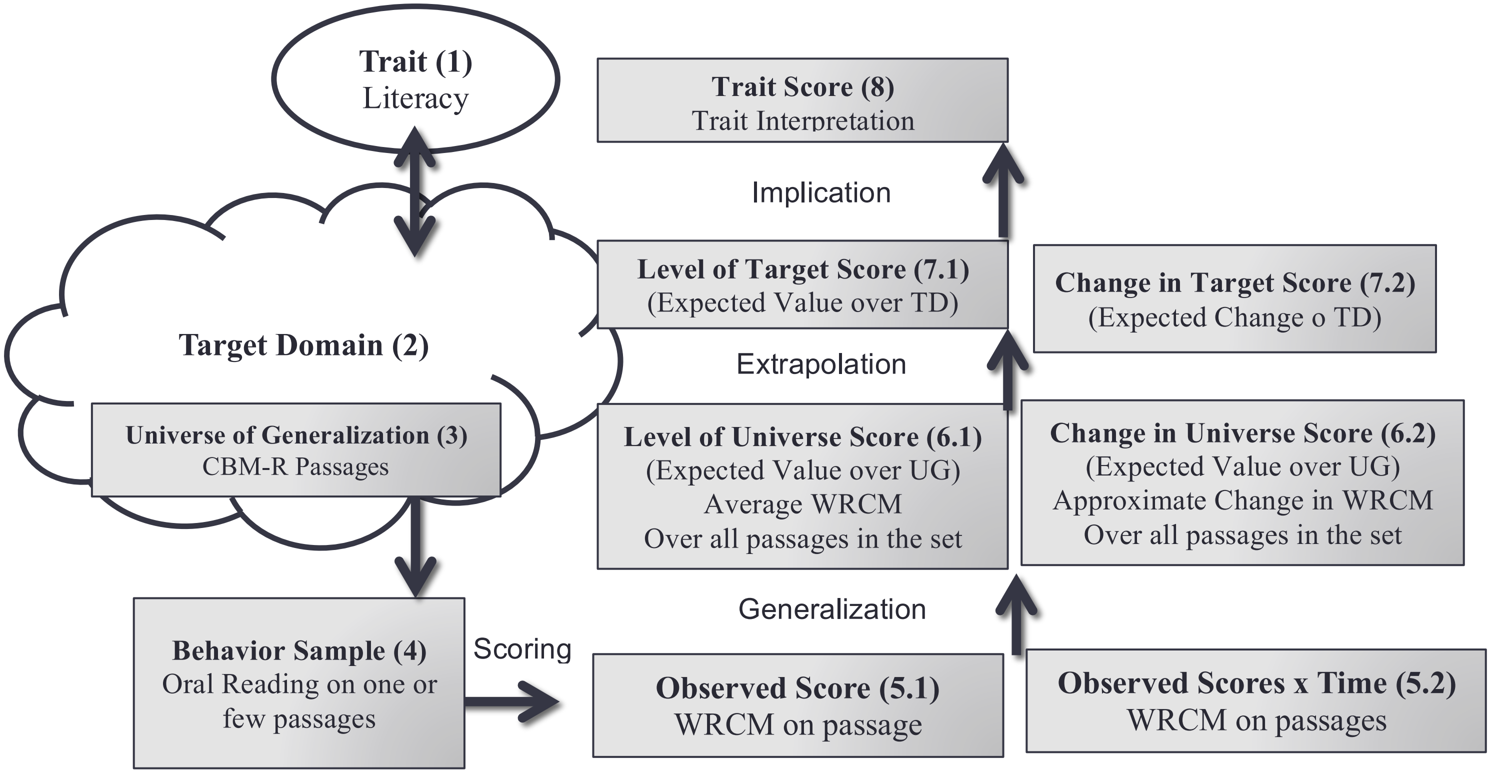

Figure 2. Example interpretative use argument for using oral reading rate as an indicator of broad reading competence. From Christ, Van Norman, and Nelson (2014).

Drawing from Generalizability theory (G-theory), the universe of generalization is conceptualized as all possible stimuli and characteristics of testing conditions (Cronbach, Linn, Brennan, & Haertel, 1997). Here the universe of generalization includes all possible grade level reading passages administered across all possible testing conditions. As Christ and colleagues (2014) mention, this universe of generalization, while still broad, greatly constricts all of the possible behaviors that may encompass the broader domain of reading achievement. To complicate matters, the manner in which CBM-R probes are developed and the conditions in which they are administered have become increasingly standardized.

When test developers and users increase the standardization of the construction of stimuli and testing conditions, the consistency or reliability of measurement typically increases. Yet, improving consistency may come at a cost of the ability to infer performance in the broader target domain (the behavior sample from the universe of generalization). Drawing from G-theory, the assumption is that the passages and the conditions in which those passages are read are representative of the broader universe of generalization (Hintze, Owen, Shaprio, & Daly, 2000; Poncy, Skinner, & Axtell, 2005). The term representative is often used in place of random because the universe of generalization is often not well defined. That is, all possible sources of variance in test performance are unknown. When considering CBM-R, the performance of a student reading out loud for one minute from a particular passage, at a particular time, in a particular setting is only a small subset of the universe of generalization. Researchers have conducted several studies using G-theory to identify factors associated with stimuli

(Christ & Ardoin, 2009; Poncy et al.) that contribute to variability in scores in the broader universe of generalization.

Moving to the right of the figure, the observed score is the numeric value given to the behavioral sample. In this case, the observed score is the WRCM on a given passage. Here an administrator listening to a student read counts each word as correct or incorrect. Therefore, the scoring inference for CBM-R is that variability in scoring (not necessarily performance) is not caused by an administrator. When educators use CBM-R as a progress monitoring tool, the scoring inference is extended by assuming that administrators will not influence the variability of scores across measurement occasions. That is, it is presumed administrators will scores passages similar to one another and consistently across time.

The next rectangle on the right hand side of Figure 2 suggests that the degree to which scores vary between specific sequences of passages is not the metric of interest with CBM-R progress monitoring data. Instead, a student's rate of improvement (ROI) or weekly increase in WRCM is assumed to generalize across all combinations of passages within a set. In other words, the ROI between passages 1, 2, 3, and 4 is assumed to be the same as if progress was monitored with passages 6, 7, 8, and 9. Further, the change in the universe score, or the ROI, is inferred to be the change in the target domain or performance score. Here, the assumption is that changes in the universe of generalization can be extrapolated to changes in broad reading achievement (the next box up on the right side). That is, if the student is improving in oral reading rate – as measured by CBM-R passages – he or she is also improving in broad reading achievement.

Finally, the interpretation of the degree of change in a trait, in this case literacy, is inferred through the degree of change in the target domain. In the context of CBM-R an initial inference is that changes in oral reading rate can be accurately detected. Among the evaluative tools to assess change in oral reading rate are visual analysis and CBM-R decision rules.

As suggested in the original interpretative argument, this project focused on a specific aspect of inferring that a change in oral reading rate is indicative of change in reading achievement and literacy. When using CBM-R to monitor instructional effects, it seems reasonable to expect that methods to detect changes in oral reading rate are in fact accurate or correct. The purpose of Study 1 was to identify relevant scenarios in which CBM-R decision rules may be applied or visual analysts may be called upon to judge changes in oral reading rate. The purpose of Study 2 was to determine which visual analytic methods and decision rules yield the most accurate decisions in those conditions. First, a literature review was conducted to determine the extent to which evaluation methods of CBM-R have already been researched.

Literature Review

The purpose of this literature review was to summarize the available evidence for the accuracy of decision rules and visual analysis to interpret CBM-R progress monitoring data. It should be noted that the purpose of this review was not to summarize evidence regarding the technical adequacy of growth estimates themselves or whether oral reading rate is in fact an adequate indicator of broad reading achievement. The purpose of this project was to evaluate the defensibility of long-standing evaluation methods. This review was guided by one question: To what degree does available evidence support the accuracy of decision rules and visual analysis to guide interpretations of CBM-R time series data?

Inclusion Criteria

To be included in the review, the work had to be an empirical study and directly address the use of CBM-R. Pieces that only summarized previous research or presented guidelines of how to conduct progress monitoring were not included. Further, the study had to evaluate the accuracy of at least one type of evaluation method – decision rules or visual analysis. In other words, some criterion of progress to measure the accuracy of the decision rule or visual analysis had to be documented. Studies that only evaluated reliability or consistency, that is the degree to which visual analysts agreed within one another, were not included. Further the criterion had to capture improvement in some capacity. In other words, a static end of year criterion was not sufficient to measure a student's ROI. Journal articles, technical reports, and dissertations were eligible for review.

Search Strategies

The search was conducted in two steps. First, a systematic ancestral review was performed. Second, a broad search of relevant databases was carried out. To perform the ancestral review, the most recent literature review on CBM progress monitoring practices was investigated (Ardoin et al., 2013). One study identified in that review purportedly evaluated the accuracy of CBM-R trend line decision rules (vanDerHedyen et al., 2005). The simulation studies referenced in the article were also reviewed. Next, the reference section of vanDerHedyen and colleagues was read. Any promising studies were identified and the same step-wise process was continued for any identified studies. In addition, the original vanDerHedyen and colleagues work was entered into google scholar. A reverse search was conducted to identify works that cited it. No eligible unique studies were found through the ancestral review. A similar ancestral review was conducted by evaluating an upcoming book chapter outlining best practices in progress monitoring (Hixson, Christ, & Bruni, 2014). One unique study was identified in the book chapter.

Online databases (Google Scholar, PsychInfo, Dissertation Abstracts, and the Research Institute on Progress Monitoring) were searched. The search term: "Curriculum-Based Measurement" was used as a baseline. The term resulted in over 600 hits in PsychInfo and Dissertation Abstracts, and over 1,000 hits in Google Scholar (after filtering for peer-referred works and dissertations). Next CBM was paired with "Progress Monitoring". This combination resulted in 89 hits in PsychInfo and over 1,000 hits in google scholar. Finally "CBM" and "Progress Monitoring" were paired with "Decision Rules" or "Visual Analysis". Only one article was found in PsychInfo using any combination of the search terms – the literature review by Ardoin and colleagues (2013). No more than 170 works were returned in google scholar for any combination of terms. No eligible unique studies were identified in any electronic databases.

A total of four studies were identified – two of which were unpublished manuscripts. Two criteria excluded several studies from the review. First, several studies investigated the reliability of evaluation methods. That is, authors calculated how consistent decisions were among similar visual analysts (e.g., Tindal, Deno, & Ysseldyke, 1983) and the consistency of decision rules were when interpreting subsets of student data (e.g., Burns, Scholin, Kosciolek, & Livingston, 2010). However, none of the excluded studies explicitly investigated the accuracy of evaluation methods.

The other factor that led to the exclusion of one study was the nature of the criterion measure. The study by vanDerHeyden and colleagues (2005) used a static end of year criterion to assess the accuracy of trend line decision rules, not a comparable metric of growth. The lack of a sound criterion to measure growth with CBM-R is in part due to the fact that one driving force behind the development of CBM was the lack of educational measures that were sensitive to change across brief periods of time (Skiba, Deno, Marston, & Wesson, 1986).

The four studies are reviewed in the next section. Three of the studies investigated the accuracy of CBM-R decision rules. The fourth study investigated the accuracy of visual analysis.

Decision Rule Studies

All of the identified works that investigated the accuracy of decision rules were part of a related line of simulation studies. In those studies a large extant dataset was first analyzed to generate model parameters. Next, true growth was specified, and observed growth was generated across different progress monitoring scenarios. Manipulated scenarios included the quality of datasets and the length and intensity of data collection schedules. Dataset quality was operationally defined as the residual variance or average level of spread of observations around a computed line of best fit (σ_ε). Highly variable data represented scenarios where administrations were not highly standardized and instruments were hastily constructed. Conversely, low residual variance represented scenarios where progress was monitored in highly standardized conditions with high quality instruments. The length of data collection schedules referred to the number of weeks progress was monitored. The intensity of data collection referred to the number of measurement occasions per week as well as the number of observations collected per occasion.

Within each study the researchers compared the recommendation of a decision rule based upon observed growth with the magnitude of true growth for each case. The researchers selected a 1.50 WRCM improvement per week goal line. If the decision rule suggested that the student was showing inadequate progress (i.e., the observed trend line was less than 1.50 WRCM or the last 3 points fell below the goal line) and the magnitude of true growth was less than 1.50 WRCM, the response was counted as correct. If the decision rule suggested that the student was making progress and the magnitude of true growth was greater than 1.50 WRCM the response was counted as correct. Conversely, when the decision rule did not suggest the student was progresses and the magnitude of true growth was greater than 1.50 WRCM or the decision rule suggested the student was progressing and the magnitude of true growth was less than 1.50 WRCM the response was counted as incorrect.

In the first study, the authors calculated the diagnostic accuracy of OLS trend lines in reference to a 1.50 WRCM goal line (Christ, Zopluoglu, & Monaghen, 2012). The researchers modeled a data collection schedule where one observation was collected per week for 3, 4, ...20 weeks. In addition, very good (σ_ε =5,) good (σ_ε =10), poor (σ_ε =15), and very poor (σ_ε =20) dataset qualities were modeled. The researchers attempted to identify the observed growth estimate where sensitivity, the proportion of cases that are in fact not benefiting from the intervention that are identified as not benefitting, and specificity, the proportion of cases that are in fact benefitting from an intervention that are identified as benefitting, were maximized and balanced across different progress monitoring durations as well as quality datasets. Stated differently, the authors identified the average weekly increase in WRCM that decreased the odds of a false positive and false negative decision equally well across different progress monitoring durations and levels of dataset quality. Across conditions, observed growth estimates that approximated 1.00 WRCM per week minimized and balanced the risk of false positive and false negative decisions when using trend line decision rules.

Aside from identifying the ideal level of observed growth to make accurate decisions, the authors also attempted to determine how quickly accurate decisions could be made using trend line decision rules. Within educational research sensitivity and specificity are ideally balanced and maximized with a lower bound of .70 for each (Hintze & Silberglitt, 2005). Using a minimum standard of .70 for sensitivity and specificity, low stakes decisions could be made after 8 weeks with very good quality datasets and after 12 weeks with good quality datasets.

The authors of the second CBM-R simulation study extended the findings of the first study by evaluating how the density (number of observations collected per occasion), frequency (the number of data collections per week) and duration of data collection schedules affected the accuracy of trend line decision rules (Christ, Zopluoglu, Monaghen, & Van Norman, 2013). The researchers sought to determine if collecting more observations each week could yield diagnostically accurate decisions in a shorter amount of time than collecting one observation per week. In other words, the researchers investigated whether low stakes decisions were feasible sooner if more data were collected each week.

As before, the researchers used a 1.50 WRCM improvement per week goal line to evaluate the diagnostic accuracy of trend line decision rules. When three observations were collected one time per week compared to one observation per week, minimal levels of sensitivity and specificity were achieved after 6 weeks of data collection, 2 weeks sooner than collecting one observation per week with a very good quality dataset. For a good quality dataset, low stakes decisions could be made after 8 weeks, 4 weeks sooner than when one observation was collected per week. In both conditions the ideal level of growth approximated 1.00 WRCM per week.

The authors of the last simulation study considered here evaluated the diagnostic accuracy of 3, 4, and 5 point decision rules (Christ, Monaghen, & Zopluoglu, 2012). For very good quality datasets, using a three point decision rule, requisite levels of sensitivity and specificity were attained after 16 weeks. Specificity often exceeded .70 for very short durations, at the expense of sensitivity. That is, the three point decision rule over identified students as benefiting from interventions at the expense of identifying students who were not. A similar pattern emerged for four and five point decision rules for very good quality datasets. Nearly 16 weeks of data were needed for four point decision rules and 20 weeks of data for five point decision rules. Making decisions with four and five point decision rules across brief periods of time came at the expense of sensitivity to an even greater degree than three point decision rules. Acceptable levels of sensitivity and specificity were not attained with good quality datasets.

As stated in Chapter 1, each simulation study evaluated the diagnostic accuracy of decision rules in a purely analytic framework. That is practitioners did not interpret the progress monitoring cases. In schools CBM-R decision rules are rarely applied indiscriminately. That is, educators are often called upon to interpret progress monitoring cases in conjunction with the recommendation of decision rules.

Visual Analysis Study

Van Norman, Nelson, Shin, and Christ (2013) investigated the accuracy of thee visual analytic methods for interpreting CBM-R progress monitoring data. In that study, the researchers created 18 fully crossed time-series graphs that presented progress monitoring cases one of three ways: (1) as a scatter plot, (2) as a scatter plot with a trend line, or (3) as a scatter plot with a 1.50 WRCM improvement per week goal line and trend line. Six graphs were available for each graphic aid. In addition, true slope was varied systematically to represent no growth (0.00 WRCM improvement per week), minimal growth (0.75 WRCM improvement per week) and substantial growth (3.00 WRCM improvement per week). Ten data points were generated for each case representing 10 weeks of data collection. Finally a random error term with a mean of 0 and standard deviation of 10 was added to each observation. Consistent with the simulation studies that level of error represented a good quality dataset.

For each progress monitoring case 52 participants indicated whether the student was not making progress, somewhat making progress, or making progress. These decisions represented slope magnitudes of 0.00, 0.75, and 3.00 WRCM respectively. Each response was coded as correct or incorrect based on the agreement of the participant’s evaluation with the pre-defined slope magnitudes. Next, the researchers used multi-level modeling to estimate the probability of a correct decision in the presence of different graphic aids across the three trend magnitudes. On average across all graphic aids and trend conditions, the average probability of a correct decision was .60. When participants used graphic aids to interpret cases that represented no growth and substantial growth the probability of a correct decision increased dramatically (.76 and .71 respectively). Regardless of graphic aids, participants struggled when interpreting cases that represented minimal growth (< .52 probability of making a correct decision). Across all trend magnitudes, participants were more likely to make a correct decision when supplied with a goal line and trend line compared to when they viewed cases without graphic aids. However, within the study, the performance of visual analysts was not directly compared to the performance of data point or trend line decision rules.

Summary

Collectively the results of the studies suggest that decision rules and visual analytic methods are capable of accurately detecting growth in oral reading rate. More specifically, it appears that the likelihood of making a correct decision with CBM-R decision rules is affected by the variability of data being interpreted, the number of observations collected, and the duration data is collected. It appears that collecting data with high quality instruments in standardized settings, multiple times per week, for over a month increases the likelihood of making a correct decision. Less is known about the circumstances in which visual analysts are likely to make correct decisions with CBM-R. This is due in part because in the single study the duration of progress monitoring and the variability of observations were not manipulated.

Within each study, the criterion for correct decisions was simulated growth. That is, none of the studies evaluated the accuracy of decision rules or visual analytic methods with actual student data. As a result, the degree to which the hypothetical data represented actual student data or conditions in which progress monitoring cases are evaluated in practice is not clear. Further, with non-simulated data, educators are not afforded the knowledge of what a student's rate of true ROI or growth is.

Last, none of the identified studies investigated the intersection of decision rules and visual analysis. That is, accuracy of interpretations based upon decision rules and visual analytic methods were not compared to one another across the same progress monitoring cases. This in direct contrast to how CBM-R progress monitoring data are actually interpreted in schools. Decision rules are rarely applied without some degree of discretion. Rather, decision rules are often applied and visual analysts also interpret the graphical displays of student data to reach a decision about instructional effects.

Conclusion

This project addressed gaps in the research literature in multiple ways. First, Study 1 provided a summary of measurement characteristics of progress monitoring cases from actual student data. This ensured that the performance of visual analysts and decision rules were investigated across relevant conditions in Study 2. Second, actual student data was evaluated and expert agreement served as the criterion measure. This provided a more ecologically valid criterion and emulated the reality of applied practice when true ROI is not known. Third, decision rules and visual analysis were investigated using the same progress monitoring cases and criterion. This allowed for the direct comparison of both types of evaluation methods.

Chapter 3

Study 1

Educators have used oral reading rate estimates from curriculum based measures as an indicator of broad reading competence for over 20 years (Wayman et al., 2007). Extensive correlational evidence suggests that oral reading rate is strongly related to and predictive of broad reading achievement (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Reschly, Busch, Betts, Deno, & Long, 2009; Shinn, Good, Knutson, Tilly & Collins, 1992; Stage & Jacobsen, 2001; Wayman et al., 2007). More in line with the original idiographic intent of CBM-R, educators also use a student's weekly rate of improvement (ROI) in oral reading rate as an indication of the effectiveness of instructional programming. Decisions to continue, modify, or terminate interventions are based upon the student's ROI across time (Deno, 1986; Deno, 1990; Stecker, et al., 2005). Further, educators use trend line and data point decision rules along with visual analysis to guide those decisions (Ardoin et al., 2013). Yet, the accuracy of time series interpretations of CBM-R progress monitoring data has not been extensively investigated. This is problematic given the potential for educators to terminate effective interventions prematurely or continue ineffective interventions unnecessarily.

CBM-R as a Progress Monitoring Tool

The evidence for CBM-R as a screening and benchmarking tool is well documented (Hintze & Silberglitt, 2005; Shinn, 1989). Yet evidence of reliability and validity should be gathered for all intended uses of a measure (AERA, APA, & NCME, 1999). That is, the psychometric evidence of CBM-R as screening and benchmarking tool does not confer evidence for its use to monitor student progress week to week (Christ et al. 2012; Fuchs, 2003). If CBM-R were a perfect index of instructional effects, instruction would be the only influence on ROI (Jenkins, Zumeta, Dupree, & Johnson, 2005). However, factors irrelevant to reading performance, including characteristics of passage sets (Ardoin & Christ, 2009; Betts, Pickart, & Heistad, 2009), the setting of data collections (Derr-Minneci, 1990) and how instructions are delivered (Colon & Kranzler, 2006) influence CBM-R estimates. In fact, CBM-R estimates can fluctuate by as much as 40 WRCM across brief periods of time (Christ & Ardoin, 2009). Most efforts to summarize the measurement characteristics and sources of construct irrelevant variance of CBM-R observations have done so within the context of single observations, not ROI (e.g., Christ & Hintze, 2007; Poncy et al., 2005).

Despite the widespread use of CBM-R to evaluate instructional effects, relatively little research has been conducted investigating the measurement characteristics of CBM-R progress monitoring data, particularly at the individual student level. Early research summarized the average level of variability of observations across time, or the standard error of the estimate (SEE; Skiba, et al., 1986). Later, studies investigated the degree of precision of different trend estimation methods (Shinn, Good, & Stein, 1989) as well as the accuracy of predictions from those methods (Good & Shinn, 1990). Researchers investigated the effect of passage type on the precision of growth estimates, quantified as the standard error of the slope (SEb; Hintze & Christ, 2004). The potential implication of passage equivalence on the consistency of growth estimates was also explored (Albano & Rodriguez, 2012; Francis, Santi, Barr, & Fletcher, 2008). However, other measurement characteristics that influence the accuracy of time-series interpretations have not received comparable attention within the CBM-R progress monitoring literature.

Characteristics of Progress Monitoring Data

Before the accuracy CBM-R decision rules or visual analytic methods are explored, the conditions in which those methods should be investigated need to be identified. Examples of those measurement characteristics include the number of observations that are collected, the variability of those observations, and the magnitude of ROI based upon those observations. If systematic differences are found, then researchers that evaluate time series interpretations might manipulate those characteristics. In addition, the authors of recent CBM-R simulation studies (Christ, Zopluoglu, Long et al., 2012; Christ, Zopluoglu, Monaghen et al., 2013), the closest investigation to the accuracy of CBM-R decision rules (Ardoin et al., 2013), generated data based upon a common extant dataset. However, the researchers did not conduct a thorough analysis of that dataset. In addition, the simulation studies only examined the influence of four variables on resulting growth estimates: (a) schedule and (b) duration of progress monitoring, (c) magnitude of growth estimates, and (d) residual. A more thorough analysis of that extant dataset, in addition to informing future studies, may help to assess the validity of the data generation models used in recent simulation studies. Other characteristics that might influence the accuracy of CBM-R decision rules and visual analysis include: outliers or aberrant values, linearity of growth patterns, the magnitude of intercept and slope estimates, and the magnitude of autocorrelation of observations.

Duration. Researchers and practitioners struggle to define the ideal duration (length of time progress should be monitored) and density (the number of observations that should be collected per occasion) of progress monitoring schedules. Those decisions must be balanced with the practical need to make instructional decisions quickly. One of the primary findings of the simulation studies was that instructional effects need time to substantiate (Christ et al., 2013). More concretely, the most influential factor on the precision of growth estimates and the accuracy of trend line and data point decision rules was the length of time progress was monitored. If practitioners collect one observation per week with a high quality passage set under highly standardized conditions, decisions can be made after approximately 12 weeks (Christ, et al., 2012). The results of another simulation study suggested that if three observations were collected once per week, decisions could be made after 8 weeks (Christ et al., 2013) Part of the impetus for the simulation studies was the lack of available evidence to guide or support some widely used decision rules and progress monitoring practices.

Ardoin and colleagues (2013) reported that the most common recommendation in the early 1980s was to collect five-to-nine CBM-R observations before applying a decision rule. Assuming one observation was collected weekly, this translated to just over a month or a little more than two months of data collection. Assuming two observations were collected per week, a decision could be made in less than a month. Between 2006-2010 a few authors began suggesting educators collect more than 20 observations before applying a decision rule (Christ, 2006; Hintze & Christ, 2004). The general recommendation was to collect two observations per week for 10 weeks. This was due, in part, to emerging research that indicated that growth estimates were highly unstable after brief periods of progress monitoring, especially if less optimal instrumentation was used for data collection (Christ, 2006; Hintze & Christ, 2004).

Christ and colleagues (2012; 2013) did not summarize the frequency of different durations of progress monitoring schedules within the extant dataset used for data generation. Summarizing the frequency of different progress monitoring schedules will indicate the nature of progress monitoring schedules initial parameters were based on. If a disproportionate number of cases lasted for extremely short durations, the parameters used for data generation may be based upon relatively unstable growth estimates.

Missing data. The impact of missing data on the accuracy of decision rules and visual analysis has not been studied extensively. Furthermore, the actual frequency of missing observations within and across cases was not summarized in the CBM-R progress monitoring simulation studies. As a result, base-rates and standards for missing-ness have not been established.

When the duration of CBM-R progress monitoring is fixed, educators can collect data less frequently without comprising the validity of growth estimates (Christ, Monaghen, Zopluoglu, & Van Norman, 2013; Jenkins and Terjeson, 2011; Jenkins Graph, & Migliorietti, 2009). For example if an educator planned to monitor a student for three months, it may be adequate to collect data every two weeks instead of every week. However, in practice, data are often missing for reasons external to planned efforts by educators. Missing data can be a result of student absences, school closings, or other logistical issues. Before inquiries are made as to why data were missing or the implications on the accuracy of time series interpretations, an important first step is to determine the actual frequency of missing data in progress monitoring cases.

In addition, the method used to derive parameters for the data generation model in the original simulation studies, linear mixed effects regression (LMER) allows for missing data in the estimation of model parameters (Long, 2012). Further, each simulated progress monitoring case within the studies contained complete data. In other words, missing data was not a consideration for authors when analyzing the extant dataset, generating progress monitoring cases, or analyzing the accuracy of decision rules from the simulated progress monitoring cases.

Variability. The SEE for any given progress monitoring case can be thought of as the typical deviation of observations around a line of best fit, or the standard deviation of residual terms. Highly variable data display a high level of spread while data with low variability display a low level of spread around the line of best fit. Christ and colleagues (2012) described datasets with average SEE values of 5, 10, 15, 20 WRCM represent "very good", "good", "poor", and "very poor" quality datasets respectively. The quality of the dataset reflects the confidence one can have that an observed score at any given point reflects the student's true oral reading rate. While desirable, "very good" quality datasets might be rare in practice. Instead, the typical SEE value one can expect using standardized conditions with a commercial quality probe set (e.g., AIMSweb, DIBELS NEXT, or FAST) is likely between 7-15 WRCM (Ardoin & Christ, 2009).

Standard error of the estimate values can also be used to estimate the standard error of the slope (SE*b*; Christ, 2006). In turn, SE*b* values are used to construct confidence intervals around individual growth estimates (Christ & Coolong-Chaffin, 2007). Christ (2006) used published estimates of SEE to estimate SE*b* to determine the stability of growth estimates for individual students. When using typical SEE values (10) for data collection schedules that are more rigorous than those typically observed in practice (two observations a week for 10 weeks), SE*b* values still approximated 0.78 WRCM per week. To illustrate the implications of this, if a student's estimated slope was 1.30 WRCM at a given week, using a 95% confidence interval, the student's true slope may be as low as -0.23 or as a high as 2.83 WRCM per week. If the slope of the goal line for that student is 1.50 WRCM per week, then the determination of whether the student was making progress may be solely a product of error. There are generally two ways to minimize SE*b*; improve the quality of the dataset (minimize SEE) or collect more observations (prolong the duration of data collection).

Within the simulation studies, SEE values were assumed to be normally distributed across progress monitoring durations (Christ et al., 2012; Christ et al., 2013). In other words, within a given data generation model the quality of a dataset, or the SD of error terms, was fixed across students and durations, so the average SEE at 5 weeks was equal to the average SEE at 20 weeks. This may explain in part why duration was so influential on the precision of growth estimates and the accuracy of decision rules relative to the quality of datasets. Yet, it is unclear if the distributional properties of SEE values do in fact behave consistently across time.

Aberrant values. Within group based designs, outliers, or aberrant values are observations that differ substantially from other scores within a distribution (Aron, Aron, & Coups, 2009). Aberrant values are often excluded from analysis because they are not representative of the population of interest. Such values are often associated with an error in data collection or data entry. Within single case design, aberrant values are similarly defined as observations that differ substantially from other observations. The difference in single case design is that all observations are drawn from the same individual (Scruggs, Mastropieri, & Casto, 1987). Across group based and single case designs, outliers are generally viewed as problematic because they can distort the interpretations of treatment effects (Nelson, Van Norman, & Christ, 2014). In the case of CBM-R progress monitoring, the intent is to sample the student's typical reading rate and not atypical low or atypical high performances.

The presence of aberrant values, particularly at the beginning and end of time series datasets, can lead to biased OLS slope estimates (Hadi & Simonoff, 1993). This is problematic given that the most common analytical tool to summarize growth within CBM-R is OLS regression. Non-parametric regression approaches such as Theil-Sen show promise to overcome difficulties associated with highly variable data (Vannest, Parker, Davis, Soares, & Smith, 2012) but have not been evaluated in the context of aberrant values.

Before exerting extensive resources to overcome the potential negative effects of aberrant values, it seems worthwhile to get a sense of how common they are. The frequency of aberrant values within CBM-R progress monitoring data has not been investigated thoroughly. In addition, an agreed upon definition of what constitutes an aberrant value has not been reached. This may be part of the reason why previous CBM-R progress monitoring simulation studies did not model aberrant values within individual progress monitoring cases.

Growth patterns. Substantial evidence suggests that CBM-R growth patterns for upper-elementary students are often curvilinear within a school year (Christ, Silberglitt, Yeo, & Cormier, 2010; Fuchs & Fuchs, 1993). As students approach fourth grade, or average 150 WRCM, growth trajectories seem to reflect a negatively accelerating curve. It also appears that growth patterns display a season effect, or growth magnitudes from Fall to Winter are greater than from Winter to Spring screening periods (Christ et al., 2010; Nese, Biancarosa, Anderson, Lai, Alonzo, & Tindal, 2012). Even when growth is measured eight times per year, as opposed to three times per year, growth patterns appear curvilinear across grades 1-8 (Nese et al., 2013). While eight observations is more in line with how CBM-R progress monitoring data are collected in practice, it is still far from the one observation per week prototypical schedule most educators follow. Further, decisions to modify instruction are typically made after relatively brief periods of time. Therefore a negatively accelerating curve near the end of the school year may have little consequence if instruction is formatively assessed within a semester.

Growth patterns were assumed to be monotonic and linear regardless of the duration of data collection within CBM-R progress monitoring simulation studies (Christ et al., 2012; Christ et al., 2013). Further fit was not compared between linear and quadratic LMER models. Consequently all of the generated cases displayed monotonic linear growth patterns.

Intercept and slope. In general, initial levels of performance, or intercept estimates, for second and third grade students average 64 (*SD* = 37) and 89 (*SD* = 40) WRCM respectively (Pearson 2012). The average growth for second grade students approximates 1.00 to 1.50 WRCM per week. The average growth for third grade students across a school year is generally between 0.84-1.20 WRCM per week. The average weekly growth is also influenced by a student’s initial performance level (Silberglitt & Hintze, 2007). Students who attain extremely high or extremely low WRCM estimates during Fall screening periods often display flatter rates of improvement compared to students who score closer to the average. In addition, students who receive special education services often display flatter rates of improvement relative to their regular education peers (Deno, Fuchs, Marston & Shin, 2001; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Graney, Missall, Martinez, & Bergstrom, 2009).

Christ and colleagues (2012; 2013) computed the variability of intercept and slope estimates from the extant dataset based on LMER analysis. Yet the degree to which slope and intercept estimates approximated a normal distribution within the extant dataset, an assumption associated with LMER, was unclear. Further the distributional properties of slope and intercept estimates were assumed to be constant across all progress monitoring durations.

Autocorrelation. Autocorrelation is the degree to which observations at one time point predicts performance at subsequent time points (Brossart, Parker, Olson, & Mahadevan, 2006). Autocorrelation can be positive or negative. Positive autocorrelation reflects a consistent upward or downward trend of observations. Negative autocorrelation is indicative of unreliable data, or observations that rise and fall inconsistently.

The presence of positive autocorrelation within single case designs can result in the misinterpretation of treatment effects (Matyas & Greenwood, 1990). Autocorrelation is often cited as a potential threat to the interpretation of treatment effects with CBM-R progress monitoring data, (Parker, Vannest, Davis, & Clemens, 2012). Parker and colleagues found lag-1 auto correlation values ranging from .15 to .56 across four CBM-R datasets. Lag-1 values greater than .20 or .25 are indicative of severe positive autocorrelation (Matyas & Greenwood, 1990). However, the presence and strength of autocorrelation within CBM-R time series data at the individual student level and across different durations has yet to be investigated. In addition, autocorrelation was not manipulated when generating progress monitoring cases across CBM-R progress monitoring simulation studies (Christ et al., 2012; Christ et al., 2013).

Purpose

The purpose of this study was twofold. First, the measurement and modeling characteristics of a large extant dataset was explored to better understand the nature of CBM-R progress monitoring. Second, that exploration was used to evaluate the validity of assumptions and findings in the previous simulation studies.

Research questions. The study was framed by the following research questions:

1. What was the typical duration (number of weeks) of progress monitoring schedules?
2. What was the frequency of cases with missing observations across durations?
3. What was the degree of variability of observations across durations?
4. What was the frequency of aberrant values (i.e., outliers) across durations?
5. To what degree did progress monitoring cases follow linear growth patterns across durations?
6. What were the distributional properties of student level intercept and slope estimates across durations?
7. What was the magnitude of autocorrelation across durations?

Methods

Participants

Christ et al., (2013) offered a brief description of the participants and materials that comprised the extant dataset used to derive model parameters for the simulation studies:

> The parameter values for the simulations were derived from empirical datasets and expert judgment. The population parameters were derived by fitting linear mixed effects regression (LMER) models to a large progress monitoring dataset of second (n=1517) and third grade students (n=1561). The demographic distribution of participants was approximately 46% girls, 53% White, 17% Black, 8% Hispanic/Latino, 6% Asian/Pacific Islander, and 2% American Indian/Alaska Native across grade-based samples. Approximately 2% of participants across grade-based samples received special education services. Data were collected in the Midwest through a federally funded project designed to provide supplemental (Tier II) reading interventions to elementary students at risk of reading difficulties. All data collectors were trained to criterion with AIMSweb training materials and assessed for administration fidelity using the Accuracy of Implementation Rating Scales (Shinn & Shinn, 2002). Inter-rater reliability data were not available, but reported estimates typically approximate or exceed .95 (Marston, 1989).

Observations were collected once per week. The preceding paragraph, along with Table 1 is the extent of the information offered regarding the extant dataset.

Procedure

First, the analyses detailed below were applied to the entire dataset as one unit. Next, relevant analyses were conducted conditioned on the duration of progress monitoring cases. For instance, cases that spanned 4 weeks comprised one group, cases that spanned 5 weeks comprised a separate group and so on.

Analyses

Seven characteristics of the extant dataset were explored. The characteristics included: (1) the duration of cases, (2) the percentage of missing data within cases, (3) the variability of observations within cases, (4) the frequency of aberrant values within cases, (5) the characteristics of growth patterns, (6) the distribution of intercept and slope estimates, and (7) the magnitude of autocorrelation within cases.

Table 1

Parameter Estimates for Progress Monitoring Data from Christ, Zopluoglu, Monaghen, & Van Norman, 2013

	N	β₀	β₁	*Var*(*b*_{0*i*})	*Var*(*b*_{1*i*})	σ_ε	Correlation between random effects ρ(*b*_{0*i*,}*b*_{1*i*})	Number of Time/Data Points Per case
		M (*SE*)	*M* (*SE*)	*M*	*M*			*range*
Datasets Grade 2	1517	31.2 (.36)	1.58 (.02)	122.06	0.43	10.35	.23	3-31
Datasets Grade 3	1561	55.69 (.45)	1.42 (.02)	214.80	0.38	11.15	.19	3-31
Combined	3078	43.10 (.37)	1.51(.02)	327.64	0.38	10.75	.07	3-31
Simulation	**9000**[a] **(300 x 30)**	**40**	**1.50**	**150**	**0.40**	**5, 10,** **15, & 20**	**.20**	**6,8,10…20**

Note. Parameters estimates derived from large datasets of progress monitoring data. *N* – number of progress monitoring cases per dataset and simulation condition.

[a] 300 batches of 30 iterations each for a total of 9000 simulated cases per condition.

Duration. The reported duration of progress monitoring cases in the extant dataset ranged from 3 weeks to 31 weeks. Yet the frequency of different progress monitoring schedules (e.g., the number of cases that lasted 3 weeks) was unknown. It was important to gain a sense of the frequency of different progress monitoring durations, particularly to investigate how different properties interacted with the length of time data were collected.

First, the average number of weeks progress was monitored for the entire sample was calculated. In addition, the median, standard deviation, kurtosis, and skew values were computed. Finally, the number of cases for each progress monitoring duration was tabulated.

Missing data. First, the number and percentage of cases with missing data in the entire dataset was calculated. If data were missing between the first recorded week and the last recorded week, the case was flagged. Within the extant dataset one observation was collected once per week. Of the cases that were flagged, the percentage of observations that were missing within the data collection schedule was calculated. After that, descriptive statistics summarizing cases with missing observations and the percentage of missing observations within those cases (mean, median, SD, kurtosis, skew) was calculated.

Variability First, SEE and SE*b* values for each progress monitoring case were computed. Descriptive statistics were then calculated for the entire sample. Next, descriptive statistics were teased apart by the duration of progress monitoring schedules.

Aberrant values. Aberrant values were defined as observations with a residual value 1.96 times greater than the average SEE for the case's comparative sample. For instance if the average SEE in a sample was 10, an observation would be considered extreme if it deviated from its predictive value for that week more than 19.60 WRCM.

First, aberrant values were identified within each progress monitoring case using the average level of SEE from the entire dataset as a referent. Every residual value for each observation within every case was calculated. If the absolute residual value was greater than or equal to 1.96 times the average SEE value of the entire dataset, the case was flagged. The overall percentage of cases with aberrant values was then calculated. Within flagged cases, the number of aberrant values was also calculated. The same series of analyses was conducted conditioned on the duration of progress monitoring schedules. In turn the average SEE used for comparison was based on progress monitoring cases of the same duration. As a result two operational definitions for aberrant values were used.

Growth patterns. Growth patterns were evaluated at the individual student level. For each progress monitoring case an OLS model assuming linear growth was calculated. Another OLS model with a time * time interaction was estimated. Violations of ordinary least squares (OLS) regression preclude statistical comparisons of model fit. However, model parameters themselves are less biased (Long, 2012). As a result, descriptive analyses of patterns of parameters across the dataset were carried out.

Intercept and slope Descriptive statistics for slope and intercept estimates were calculated across the entire dataset. The same analyses were carried out conditioned on the duration of progress monitoring cases.

Autocorrelation. The degree of lag-1 autocorrelation was computed within each progress monitoring case. Descriptive statistics for autocorrelation estimates were computed across the entire group and teased apart by duration.

Results

Duration

The mean duration of progress monitoring cases was 17.94 weeks (*SD* = 8.78; see Table 2) while the median was 16 weeks. The range was 3-34 weeks (see Figure 3), which conflicts with descriptive data presented in the initial simulation studies (see Table 1). The disagreement between the current results and the summary presented in CBM-R progress monitoring simulation studies could be due to the authors using the descriptors "time points" and "observations" interchangeably. That is, the authors may have assumed that the number of observations within a progress monitoring case signified the number of weeks data were collected without considering that data may be missing. Nevertheless, visual analysis of Figure 3 suggests a multi-modal distribution with three duration groups. A substantial proportion of cases lasted between 8 and 14 weeks. A separate sizable portion of cases lasted between 19-21 weeks. Finally, a number of cases lasted a little over 30 weeks. As a result, measures of central tendency presented in Table 2 should be interpreted with caution.

Table 2

Descriptive Statistics of Characteristics of Progress Monitoring Cases (n = 3,078)

	Descriptive Statistic				
Characteristic	Mean	Median	SD	Skew	Kurtosis
Duration (Weeks)	17.94	16.00	8.78	0.27	-1.18
% Observations Missing	18.15	16.00	13.34	0.96	1.26
Intercept Estimate	53.14	52.94	19.35	0.09	-0.59
Slope Estimate	1.99	1.61	1.77	1.89	10.04
SEE	9.70	9.60	3.56	5.20	122.51
SE*b*	0.75	0.50	0.80	4.07	26.40
Lag-1 Autocorrelation	-0.08	-0.07	0.29	-0.19	-0.40

Note. Intercept and slope estimates were based on monotonic ordinary least squares regression models fitted to each case. % Observations Missing – the percentage of observations within each progress monitoring case that were missing, SEE – standard error of the estimate, SE*b* – standard error of the slope.

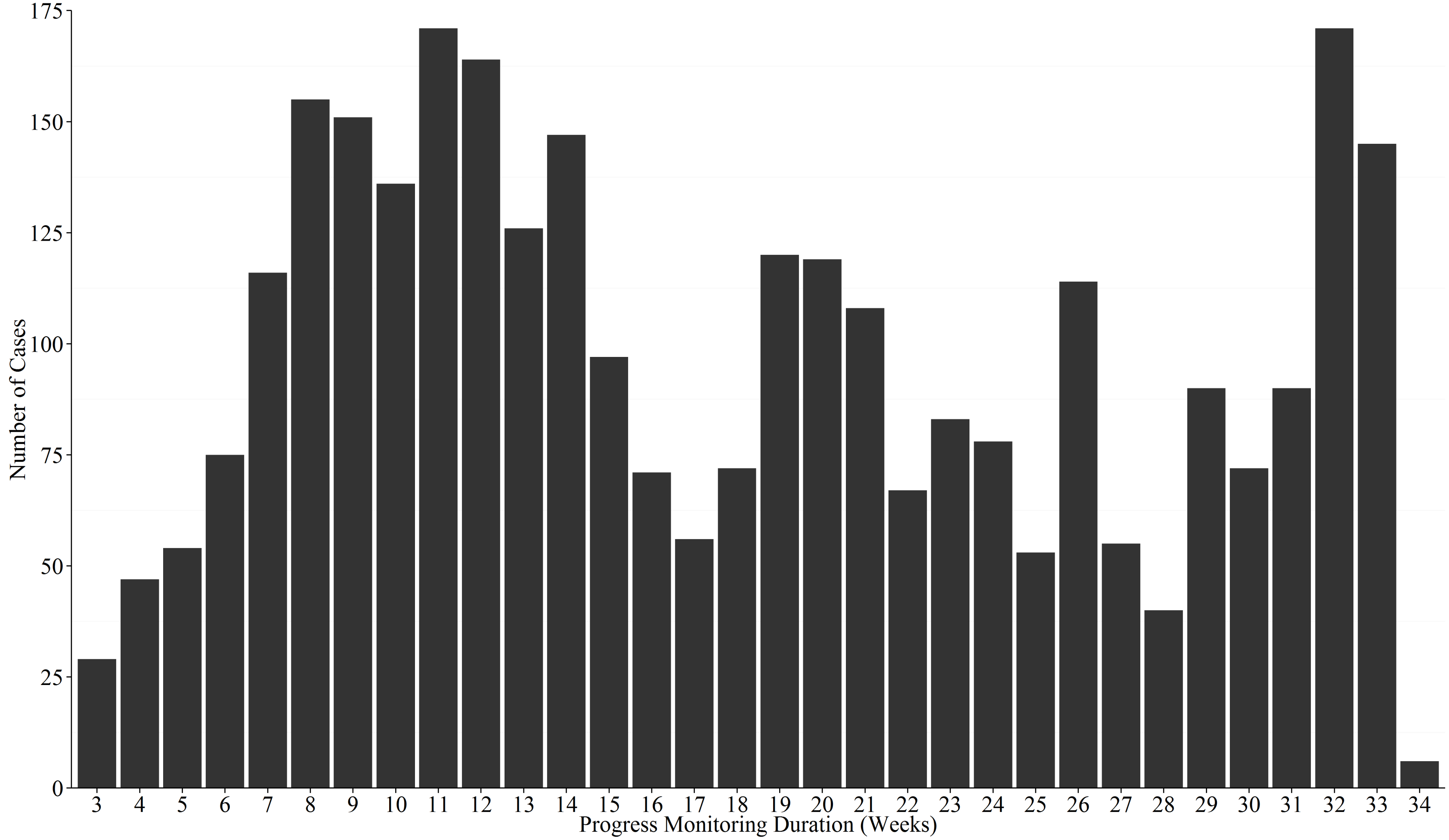

Figure 2. Number of cases for each progress monitoring duration schedule measured in weeks (n = 3,078).

Missing Data

Across the entire dataset 85% of progress monitoring cases displayed missing-ness (n = 2,635). The mean percentage of missing observations within a progress monitoring case was 18.15 (*SD* = 13.34; see Table 2). Figure 4 summarizes the proportion of observations missing within progress monitoring cases conditioned on the duration of data collection schedules. Black dots in the figure represent the mean percentage of missing observations within progress monitoring cases for that duration. While the pattern appeared to be slightly curvilinear, as the duration of progress monitoring schedules increased the percentage of missing observations tended to increase as well. More specifically, for cases that lasted longer than 10 weeks, very few, if any, progress monitoring cases possessed complete data.

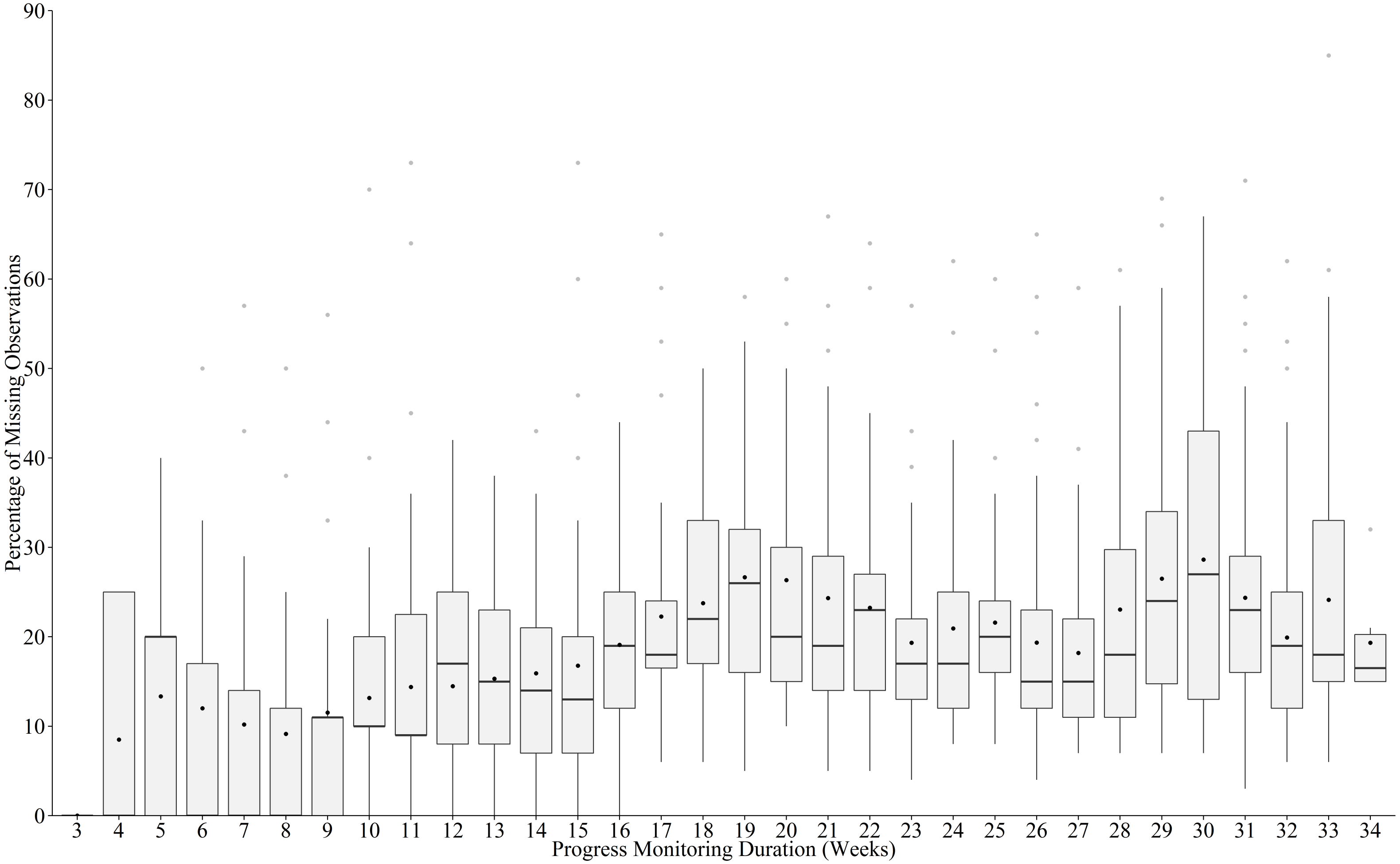

Figure 3. Box and whisker plots summarizing the percentage of missing observations within progress monitoring cases (n=3,078) conditioned on the duration of data collection schedules.

Variability

The mean SEE value across all progress monitoring durations was 9.70 WRCM (*SD* = 3.76; see Table 2) and the median value was 9.60 WRCM. The distribution of SEE values within the entire dataset was severely positively skewed (z = 5.20) and leptokurtic (z = 122.51). Christ and colleagues (2012) defined SEE values of 5, 10, 15, and 20 as very good, good, poor, and very poor respectively. The mean SEE value coincided with a good quality dataset and converged with the SEE value presented in the initial descriptive results of the simulation studies (see Table 1). Figure 5 summarizes the distribution of SEE values conditioned on the duration of progress monitoring schedules. An evaluation of the figure indicated that after 5-6 weeks SEE values seemed to stabilize. In addition, the inter-quartile range seemed to be relatively similar across progress monitoring durations, meaning that the variability of SEE values did not differ substantially as a function of time. For instance, at week seven, fourteen, and twenty one, interquartile ranges were 6-12, 8-11, and 7-11 WRCM respectively. An evaluation of the figure indicated that some durations did seem to have several outliers, or cases with abnormally high SEE values, particularly weeks 8, 9, and 10.

Whereas the SEE is an index of precision of point estimates, SE*b* is a measure of precision of growth estimates. The mean Se*b* value for the entire dataset was 0.75 (*SD* = 0.80; see Table 2) while the median was 0.50 WRCM per week. Similar to SEE values, the distribution of Se*b* values across the entire dataset were positively skewed (z = 4.07) and leptokurtic (z = 26.40). Figure 6 presents Se*b* values conditioned on the duration of progress monitoring cases. Standard error of the slope values were extremely high when the duration of progress monitoring schedules was low. This was expected since the number of observations and the duration of progress monitoring factor into the calculation of Se*b.* Values seemed to approach 1.00 after 10-11 weeks of data collection. After about 15 weeks when values approximated 0.50 WRCM per week, a point of diminishing returns to decrease SE*b* values emerged.

Aberrant Values

As stated before, aberrant values were operationally defined in two ways. First an observation within a progress monitoring case was considered aberrant if it's absolute distance from an OLS trend line (or absolute residual) was greater than 1.96 times the mean SEE value of the entire dataset. An observation was also flagged as an aberrant value within a progress monitoring case if the absolute distance of the observation from the ordinary least squares regression line (or absolute residual) was greater than 1.96 times the mean SEE value of all progress monitoring cases that lasted the same duration as that case.

Using the entire dataset as a referent, the critical value was approximately 19 WRCM (1.96 x 9.70 = 19.01). Using 19.01 WRCM as a cutoff, 1,829 (59.42%) progress monitoring cases did not have an aberrant value (see Table 3). Over 20% (n = 701) of cases had at least one aberrant value. Using each case's duration group as a referent, 1,888 (61.34%) progress monitoring cases did not have an aberrant value. Highly similar results were observed regarding the proportion of cases with aberrant values using either operational definition. Table 4 displays the proportion of cases within each duration group with different levels of aberrant values using the second operational definition. A review of Table 4 indicated that as the duration of progress monitoring cases increased, the frequency of aberrant values within those cases increased as well. After 13 weeks of data collection, nearly half of all progress monitoring cases had at least one aberrant value.

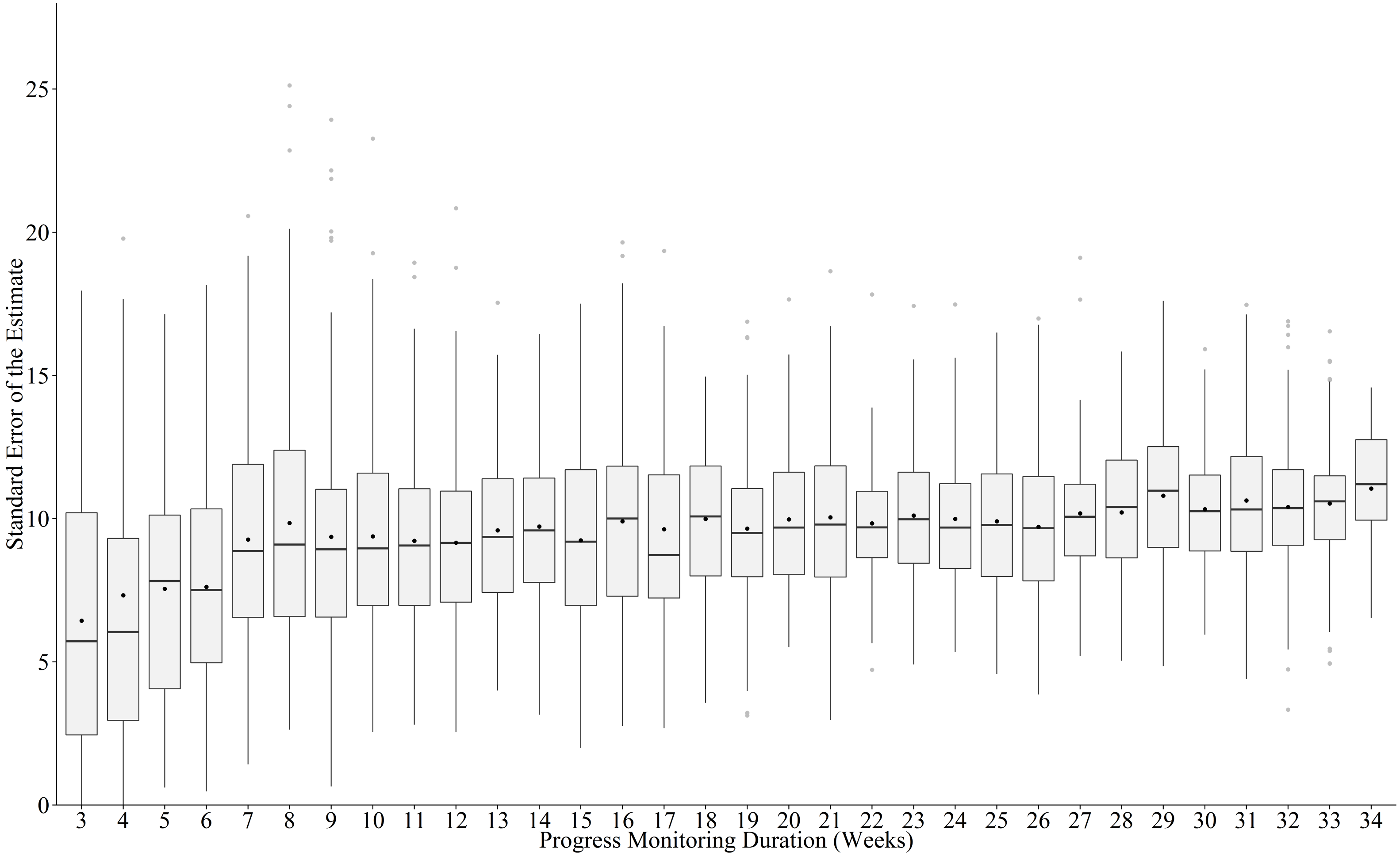

Figure 4. Box and whiskers plot summarizing standard error of the estimate values for progress monitoring cases (n = 3,078) conditioned on the duration of data collection schedules.

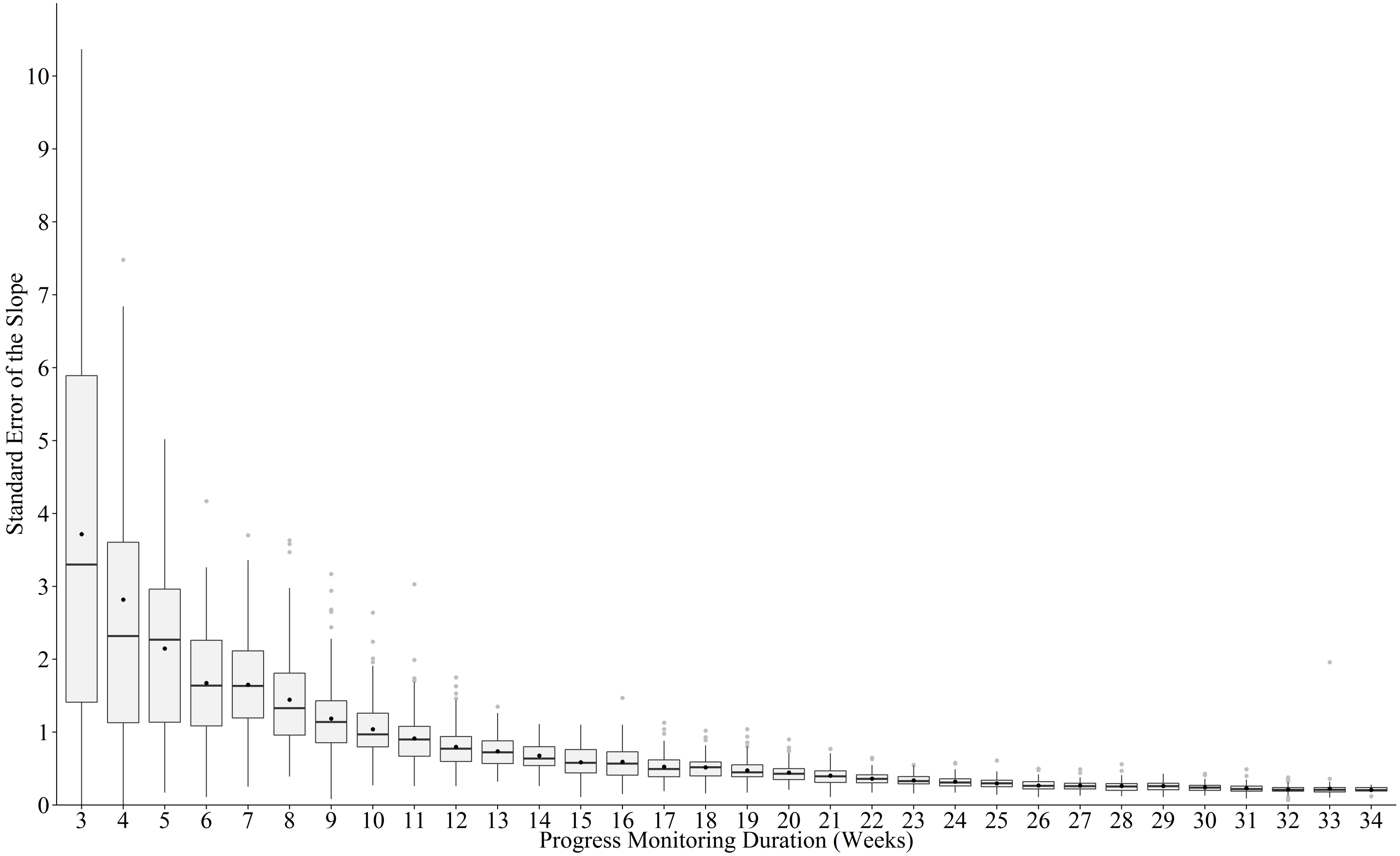

Figure 5. Box and whiskers plot summarizing standard error of the slope values for progress monitoring cases (n=3,078) conditioned on the duration of data collection schedules.

The number of cases that fell into four categories were calculated to determine whether aberrant values were a unique characteristic of progress monitoring cases, and not confounded with the magnitude of SEE. Those categories were: (1) above average variability with an aberrant value present, (2) above average variability without an aberrant value, (3) below average variability with an aberrant value, and (4) below average variability without an aberrant value. Cases were defined as having above average variability if their SEE value was greater than the average SEE value of the entire dataset (9.70 WRCM). Observations within progress monitoring cases were considered extreme using the first operational definition explained previously. Table 5 presents the number of cases satisfying each condition. An evaluation of the table indicated that cases with an extreme aberrant value and below average variability were extremely rare (approximately 1% of all cases in the dataset). Further, almost all cases that had an aberrant value also had above average variability. Therefore, it appears that the operational definition proposed to identify aberrant values within this study is likely confounded with SEE.

Growth Patterns

As stated before, an ordinary least squares (OLS) regression line depicting a negatively accelerating curve (or quadratic model) was fitted to each case.

$$\widehat{WRCM} = b_0 + b_1(Week) + b_2(Week \quad * Week)$$

Where WRCM is words read correct per minute. If a monotonic linear growth pattern is adequate to describe the pattern of observations within cases, the b_2 term in Equation 1 should be close to zero. If however growth patterns are not linear, and decelerate as time increases, parameter values should be substantially negative. To protect for over-fitting, only cases with four or more observations were considered (*n* = 2,937).

The mean value for the interaction term across all progress monitoring cases was -0.14 WRCM per week (*SD* = 0.53). Teasing the results apart by the duration of cases, schedules that lasted less than 13-14 weeks showed a high level of variability in the parameter estimate (see Figure 7). However, as time increased beyond 20 weeks, the majority of cases had interaction terms near 0 WRCM per week. A substantial proportion of cases that lasted less than 12 weeks had negative interaction terms, representing a negative accelerating curve. As a result, it appeared that negative accelerating growth patterns became less common as the duration of progress monitoring schedules increased.

Table 3

Summary of the Number and Proportion of Progress Monitoring Cases with Aberrant Values

	Entire Dataset Referent		Duration Group Referent	
Aberrant Observations in Case	Number of Cases	%	Number of Cases	%
0	1829	59.42	1888	61.34
1	701	22.77	751	24.40
2	301	9.78	277	9.00
3	151	4.91	106	3.44
4+	96	3.12	56	1.82

Note. An observation was considered aberrant within a progress monitoring case if it deviated from the predicted value of an ordinary least squares trend line more than a specified critical value. The critical value for the Entire Dataset Referent was 1.96 x the mean standard error of the estimate (SEE) of all progress monitoring cases in the dataset (19.01 WRCM; *n* = 3,078). The critical value for the Duration Group Referent was 1.96 x the mean SEE of all cases that lasted the same duration as the case being evaluated.

Table 4

Frequency and Proportion of Cases with Aberrant Values Conditioned on Duration

		Aberrant Values							
		0		1		2+		Total	
Duration (Weeks)	n	Cases	%	Cases	%	Cases	%	Cases	%
3	29	28	96.55	1	3.45	0	0	1	3.45
4	47	39	82.98	8	17.02	0	0	8	17.02
5	54	49	90.74	3	5.56	2	3.70	5	9.26
6	75	71	94.67	4	5.33	0	0	4	5.33
7	116	105	90.52	10	8.62	1	0.86	11	9.48
8	155	112	72.26	32	20.65	11	7.10	43	27.74
9	151	112	74.17	31	20.53	7	4.64	38	25.17
10	136	104	76.47	20	14.71	12	8.82	32	23.53
12	171	122	71.35	31	18.13	18	10.53	49	28.65
13	164	120	73.17	33	20.12	11	6.71	44	26.83
13	126	83	65.87	34	26.98	9	7.14	43	34.13
14	147	83	56.46	48	32.65	16	10.88	64	43.54
15	97	59	60.82	27	27.84	11	11.34	38	39.18
16	71	39	54.93	15	21.13	17	23.94	32	45.07
17	56	36	64.29	16	28.57	4	7.14	20	35.71
18	72	34	47.22	27	37.50	11	15.28	38	52.78
19	120	77	64.17	26	21.67	17	14.17	43	35.83
20	119	71	59.66	27	22.69	21	17.65	48	40.34
21	108	53	49.07	31	28.70	24	22.22	55	50.93
22	67	41	61.19	15	22.39	11	16.42	26	38.81
23	83	45	54.22	23	27.71	15	18.07	38	45.78
24	78	34	43.59	28	35.90	16	20.51	44	56.41
25	53	25	47.17	13	24.53	15	28.30	28	52.83
26	114	57	50.00	29	25.44	28	24.56	57	50.00
27	55	26	47.27	13	23.64	16	29.09	29	52.73
28	40	19	47.50	14	35.00	7	17.50	21	52.50
29	90	29	32.22	31	34.44	30	33.33	61	67.78
30	72	33	45.83	23	31.94	16	22.22	39	54.17
31	90	40	44.44	27	30.00	23	25.56	50	55.56
32	171	75	43.86	64	37.43	32	18.71	96	56.14
33	145	63	43.45	47	32.41	35	24.14	82	56.55
34	6	4	66.67	2	33.33	0	0	2	33.33

Note. An observation was considered aberrant within a progress monitoring case if it deviated from the predicted value of an ordinary least squares regression line more than a specified critical value. The critical value to determine if an observation was aberrant was 1.96 x the mean standard error of the estimate of all cases of the same duration.

Table 5

Frequency and Proportion of Cases with and without Aberrant Values Conditioned on the Magnitude of Variability of Observations.

		Standard Error of the Estimate	
		Below 9.70 WRCM	Above 9.70 WRCM
Aberrant Observations	0	1269 (41.22%)	619 (20.11%)
	1+	57 (01.85%)	1133 (36.82%)

Note. An observation was considered aberrant if it deviated from the predicted value of an ordinary least squares regression line more than 19.01 words read correct per minute within a case.

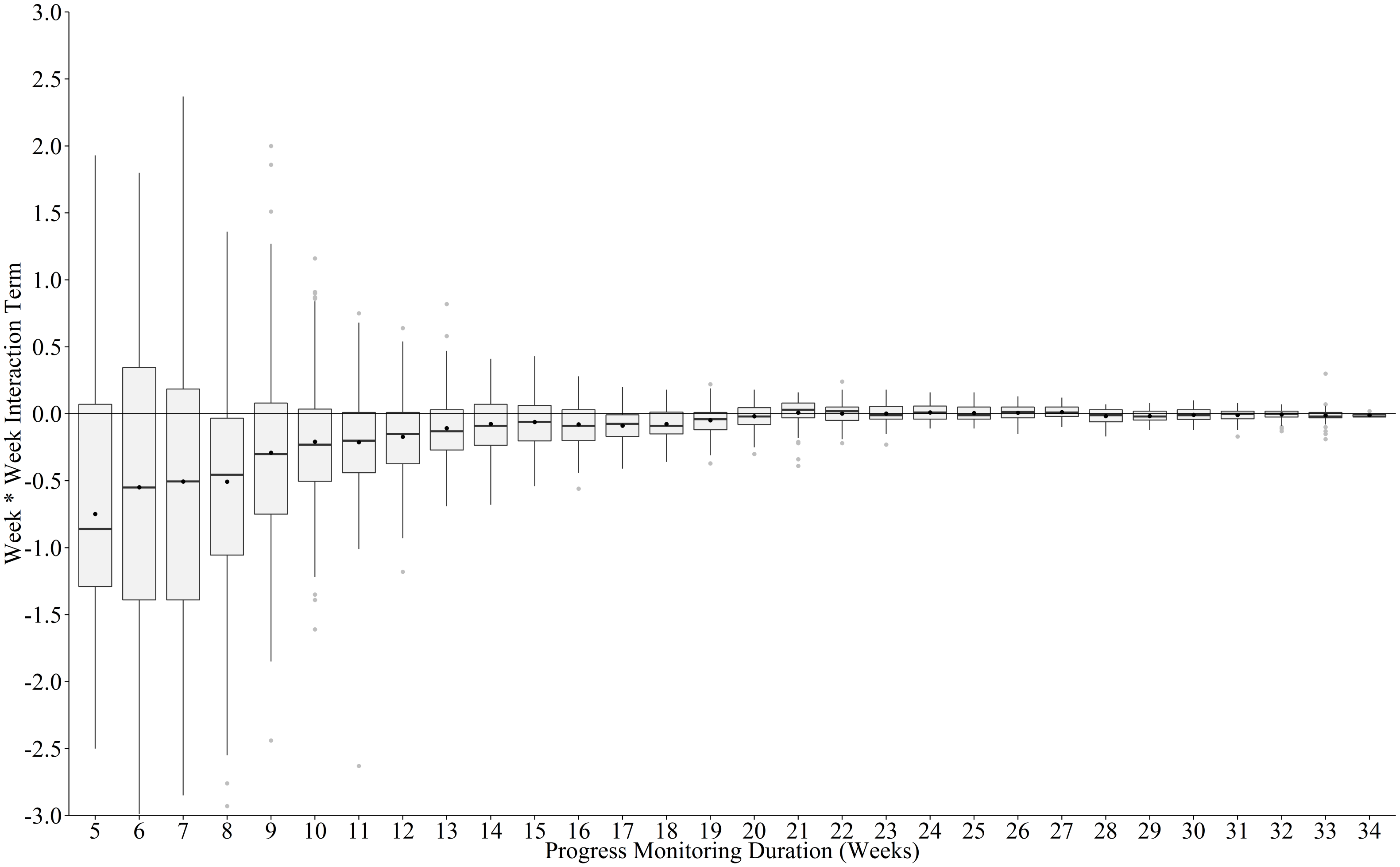

Figure 6. Box and whisker plots summarizing unstandardized week * week regression terms for progress monitoring cases conditioned on the duration of data collection.

Intercept and Slope

The mean intercept estimate across all progress monitoring cases was 53.14 WRCM (*SD* = 19.35; see Table 2). While the median intercept estimate was 52.94 WRCM. These values were slightly higher than the estimates reported in the initial simulation studies (see Table 1). The distribution of intercept estimates appeared relatively normal. The average intercept estimate increased as progress monitoring duration increased through week eight (see Figure 8). After eight weeks the average intercept estimate appeared to decrease slightly as the duration of progress monitoring increased. At week three the average intercept estimate was slightly less than 60 WRCM. By week twenty-four the average intercept was near 45 WRCM. The average intercept value seemed to fall within 10 WRCM of that value for the remaining durations.

The mean monotonic linear slope estimate across all progress monitoring cases was 1.99 WRCM per week (*SD* = 1.77; see Table 2). While the median slope estimate was 1.61 WRCM per week. Both values were higher than what was reported in the initial simulation studies (see Table 1). The distribution of slope estimates across the entire extant dataset was slightly positively skewed (z = 1.89). Distributions of slope estimates conditioned on the duration of progress monitoring schedules are presented in Figure 9. The average slope estimate for each duration fluctuated substantially until approximately 14 weeks. For instance the average slope estimate at week three was 5.00 WRCM per week and at week five the average slope estimate was less than 2.00 WRCM per week. The variability of slope estimates decreased markedly as the duration of progress monitoring cases increased through 20 weeks. The inter-quartile range for slope estimates at week three was between 0 and 10.50 WRCM per week. At week seven the inter-quartile range was between 1.50 and 4.50 WRCM per week. By week 20 the inter-quartile range decreased to between 1.50 and 2.00 WRCM per week.

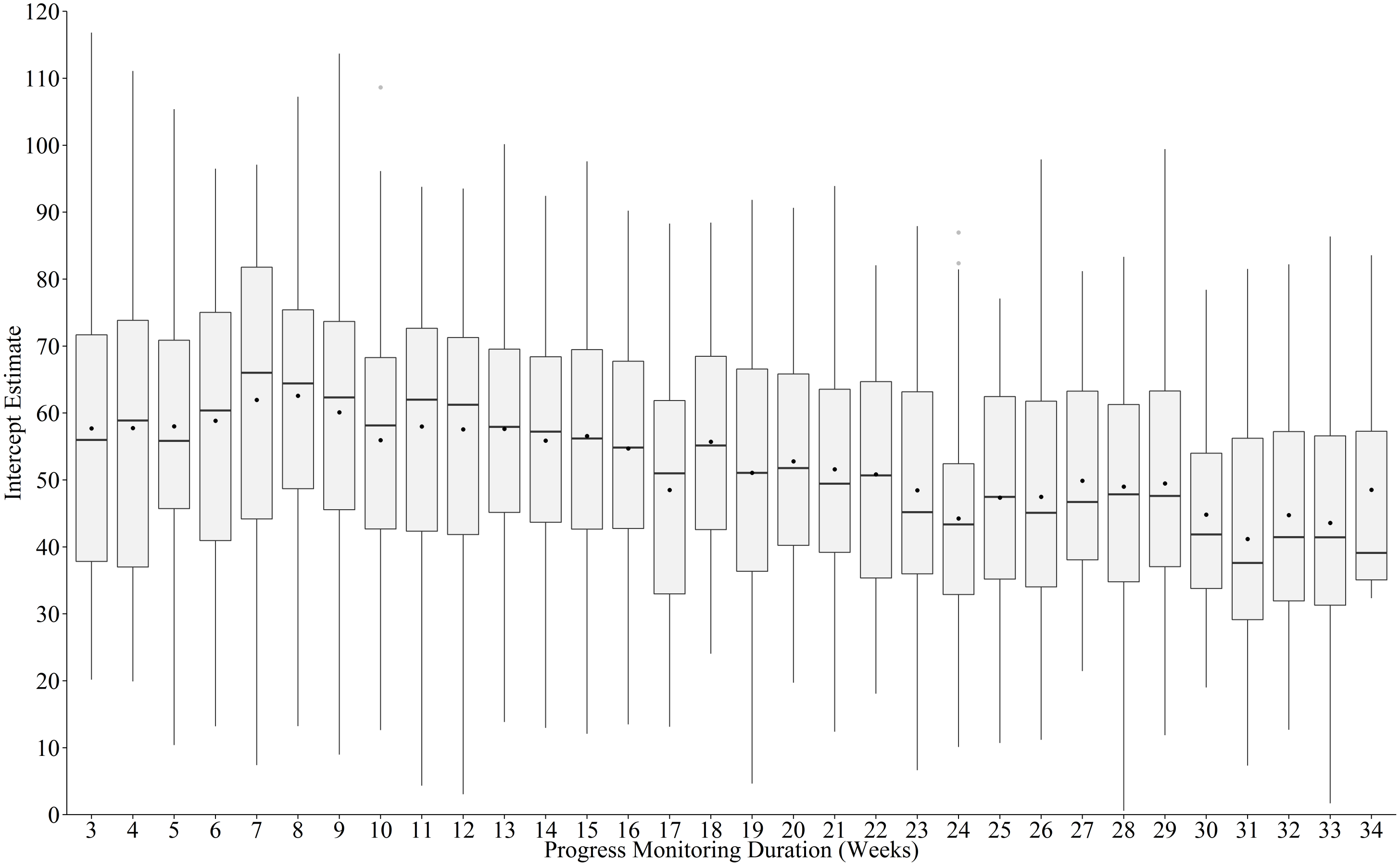

Figure 7. Box and whisker plots summarizing intercept estimates (predicted words read correct per minute score at the beginning of data collection) for progress monitoring cases (n=3,078) conditioned on the duration of data collection schedules.

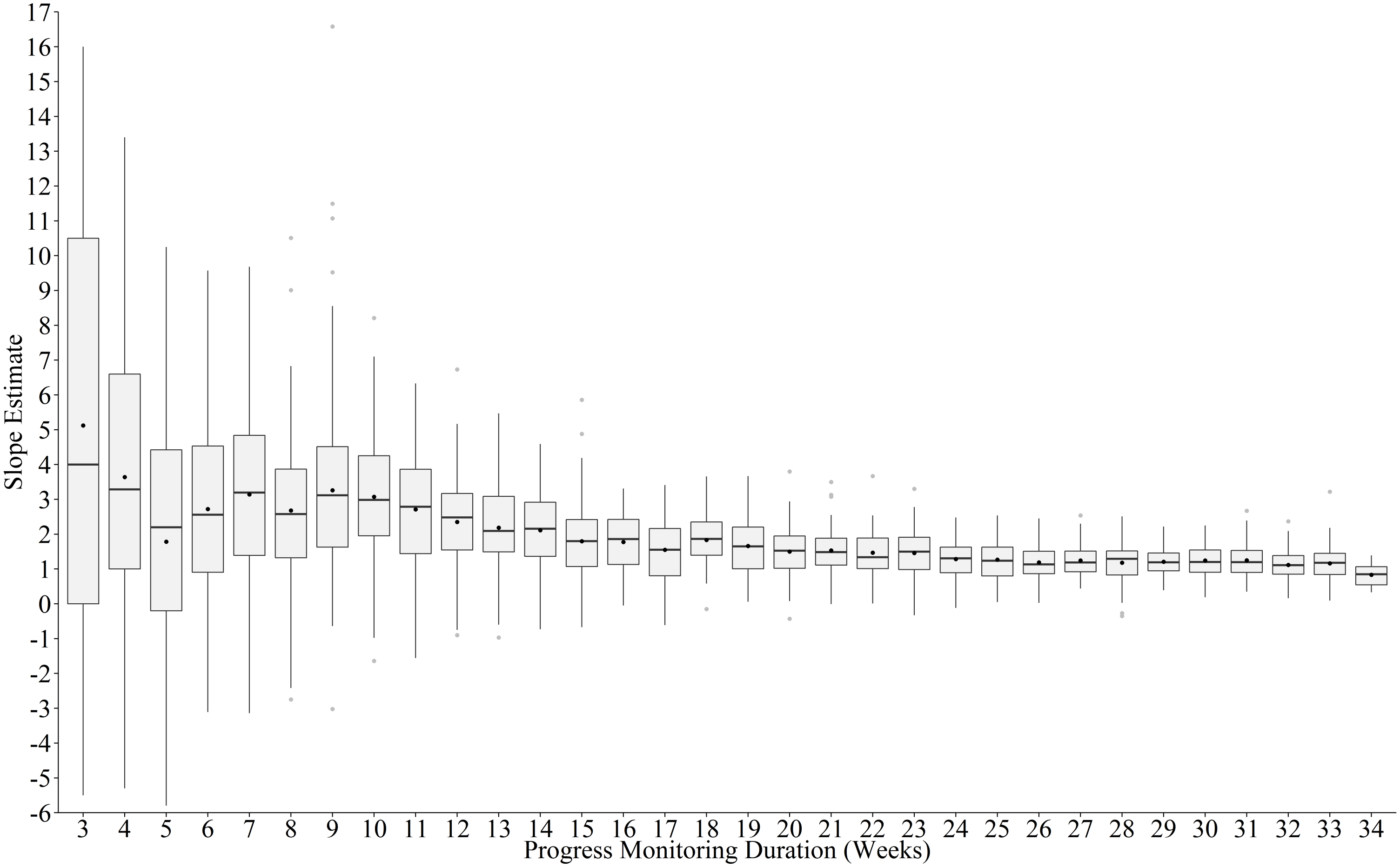

Figure 8. Box and whisker plots summarizing slope estimates (average weekly increase in words read correct per minute) of progress monitoring cases (n = 3,078) conditioned on the duration of progress monitoring schedules.

Autocorrelation

Across all progress monitoring cases, the average degree of lag-1 autocorrelation was -0.08 (*SD* = 0.29; see Table 2). The median degree of lag-1 autocorrelation was -0.07. The distribution of values was relatively normal. Within single case time series data, negative autocorrelation values are not as troublesome as positive autocorrelation values. Generally, a negative lag-1 autocorrelation value is indicative of instability of observations. That is, observations tend to fall above and below a line of best fit inconsistently. Conversely, positive autocorrelations demonstrate patterns where observations are successively increasing or decreasing. Positive auto correlation values, particularly above .12 often result in the overestimation of treatment effects (Parker et al., 2012). Figure 10 delineates the distribution of lag-1 autocorrelation estimates by the duration of progress monitoring schedules. Generally, progress monitoring schedules lasting less than 15 weeks were likely to be characterized as having negative lag-1 autocorrelation. This is in line with other characteristics of the extant dataset. That is, the shorter the duration of progress monitoring the more likely data were unreliable. Positive autocorrelation does not appear to be a major concern until schedules begin to span over 25-26 weeks. Even then, mean estimates hover near zero.

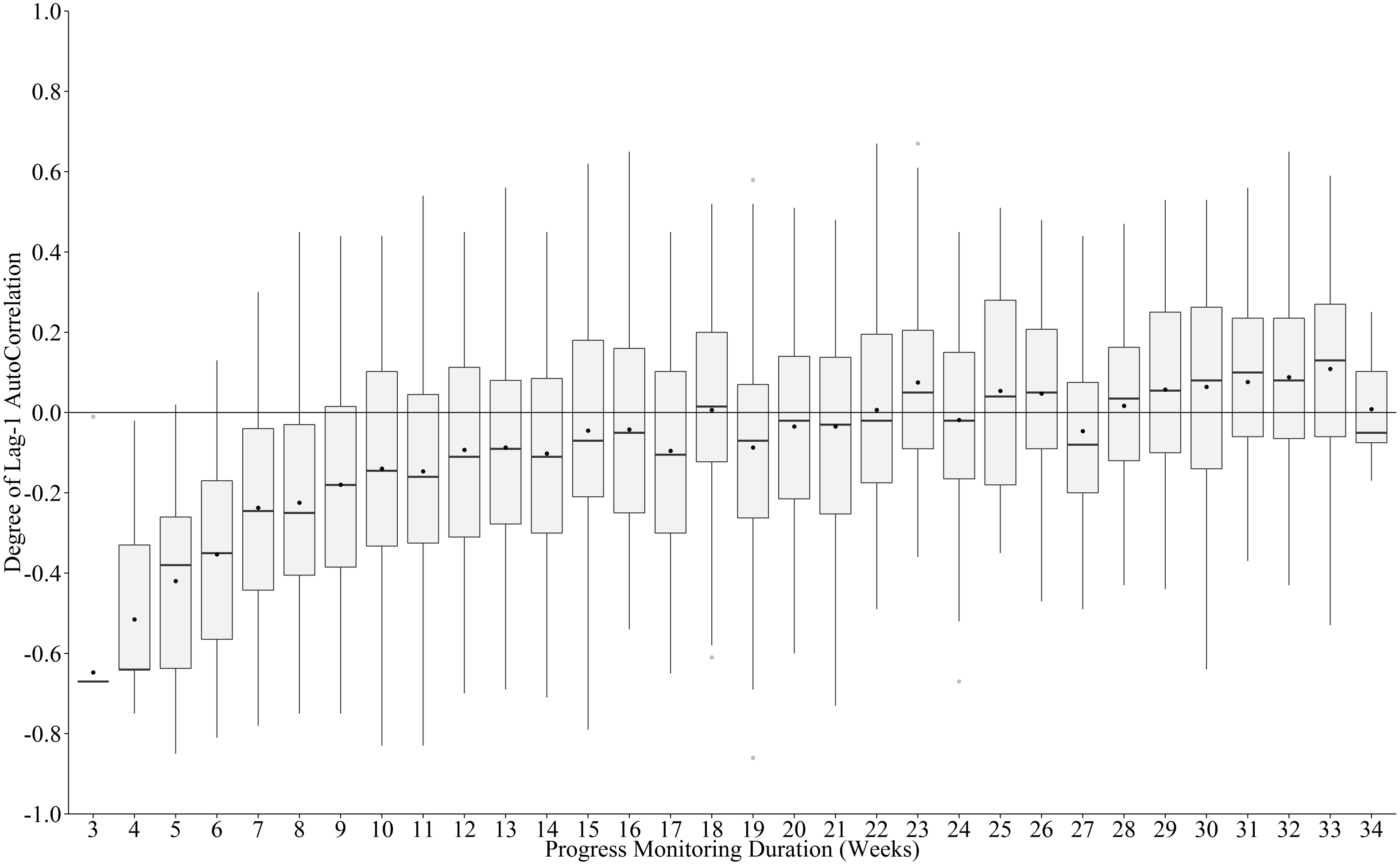

Figure 9. Box and whisker plots summarizing the degree of lag-1 autocorrelation in progress monitoring cases (n = 3,078) conditioned on the duration of data collection schedules.

Discussion

There were two broad purposes of this study. First, the measurement and modeling characteristics of a large extant dataset were explored to better understand the nature of CBM-R progress monitoring. Second, that exploration was used to evaluate the validity of assumptions and findings in the previous simulation studies. Seven measurement characteristics within the large extant dataset were summarized: (a) the duration or number of weeks progress was monitored, (b) the missing-ness of data within cases, (c) the variability of observations, (d) the frequency of outliers or aberrant observations, (e) the linearity of individual growth patterns, (f) the nature of intercept and slope estimates, and (g) the magnitude of autocorrelation within cases. The following is organized to first describe the characteristics of progress monitoring data and then evaluate the validity of the models and findings of previous simulation studies.

Characteristics of Progress Monitoring Data

Duration. The duration of progress monitoring cases within the extant dataset varied considerably. A substantial proportion of cases lasted longer than 30 weeks. This translated to over seven months of data collection. If CBM-R is used to guide instructional decisions within a school year, it is hard to justify monitoring the effectiveness of an intervention for over seven months. A sizable portion of progress monitoring cases lasted between 8 and 14 weeks. That range of durations coincides with recommendations in the research literature (Ardoin et al., 2013). However, one has to tread cautiously when inferring why different progress monitoring cases spanned different durations within this extant dataset. The reason for termination of progress monitoring is not recorded in the dataset. It could be that students who were demonstrating adequate growth were exited from the program. Alternatively, students who showed a lack of improvement may have been qualified for special education services and began receiving more intensive interventions.

Missing data. Missing data was widespread throughout progress monitoring cases in the extant dataset. After 10 weeks, almost all progress monitoring cases were missing at least one observation. Further, even progress monitoring cases that spanned four to six weeks were missing data. Recall that this extant dataset was not the result of a research study. As a result it seems that missing data reflects a reality of practice. As a consequence, within schools, it is extremely unlikely that educators would have the luxury to apply decision rules or visually analyze complete time series data when making decisions about student progress. Stated differently, missing data might be the norm. Researchers should account for missing-ness in their research on decision rules.

Variability. The average variability of observations, or SEE, across the entire dataset coincided with values reported in the initial simulation studies. In addition, the simulation studies assumed that SEE values were constant across time. Extremely brief data collection schedules aside, such as three to four weeks, SEE values did seem to stabilize across progress monitoring durations.

While the average SEE seemed to stabilize to around 8.00 WRCM across durations, there was still variability at each time point. The interquartile range for several durations spanned from 5-12 WRCM. As a result, SEE magnitude may still be a worthwhile factor to manipulate when investigating the accuracy of decision rules and visual analytic methods.

Aberrant values. Within this study the identification of aberrant values, or outliers, was confounded with the variability of observations within progress monitoring cases. That is, there were very few cases (< 1%) that had below average SEE values and possessed an aberrant observation. Notwithstanding, some progress monitoring cases had observations that deviated more than 40 or 50 WRCM from a line of best fit.

Clearly, the issue of outliers or aberrant values warrants concern. However, the operational definitions used in this study had limitations. Specifically, aberrant observations were identified based upon measures of variability that included them in their estimation. As a result, if a case did have an aberrant observation, that observation may have inflated the SEE value for the case and increased the threshold needed to identify the observation as aberrant. This could be why that almost all cases with an aberrant observation had SEE values that were above average. Researchers should collaborate and pull from other fields such as computer science, economics, or actuarial science to derive robust methods to identify outliers that are not confounded with the variability of other observations in the time series.

Growth patterns. Contrary to previous research, observations that spanned longer durations did not appear to demonstrate non-linear growth patterns. In fact observations that spanned shorter durations, particularly less than 10-12 weeks, had larger interaction terms than cases where progress was monitored the entire school year. Several things may have caused the differences in findings. First, previous studies that investigated the linearity of growth patterns compared model fit at the group level. That is, linear and curvilinear models were fit to an entire extant dataset (not individual students). Here, two separate models were estimated for each student. The latter approach, that is the approach taken in this study, reflects how progress monitoring data are interpreted in schools – at the individual student level. Second, the fact that short durations showed substantial negative interaction terms may be a result of model over-fitting. Recall, that even at 5-10 weeks a large number of progress monitoring cases had missing data. As a result, although progress was monitored for 5-10 weeks models may have been estimated only using four observations. Third, students that were monitored for longer durations may have systematically differed from students who were monitored for shorter durations. To clarify, students that were monitored for over 20 weeks may have been kept in the program because they did not show an immediate jump in performance like students who were only monitored for 4-10 weeks.

Intercept and slope. Both intercept and slope estimates were slightly higher than values reported in the initial simulation studies. This is likely due to the difference in analytical methods used to summarize the extant dataset. Estimates reported in Table 1 were in essence fixed effects, or group level averages, after partitioning serial dependency (or within subjects error) of observations. Whereas the intercept and slope estimates reported were averages based upon individual OLS estimates, not controlling for serial dependency in observations. Slight differences in LMER and OLS parameter estimates are common (Long, 2012). As a result, the more pressing concern for the purposes of this study and the appropriateness of the data generation model in the simulation studies are distributional properties of intercept and slope estimates.

Overall, the distributions of intercept and slope estimates appeared relatively normal. That is neither were so skewed as to likely present problems for LMER analysis. The most interesting finding was the difference in variability of slope estimates as a function of the duration of progress monitoring. Cases that spanned less than 8 weeks showed considerably more variation in slope estimates relative to cases that spanned over 20 weeks. Drawing from the distribution of SE*b* values it seems likely that duration is highly influential on the precision of growth estimates. Alternatively, as discussed previously, it may be that students who were exited from the program quickly showed substantially higher growth early on-indicating that the intervention was effective, or extremely low growth-indicating they required more intensive services.

Autocorrelation. The presence of autocorrelation and how to address it has been a contentious issue for over 30 years within single case design (Huitema, 1986). The issue of autocorrelation has only begun to receive similar attention within the CBM-R progress monitoring research literature. Recall, positive lag-1 autocorrelation is generally a serious threat to misinterpreting treatment effects. Based upon the results of this study, positive lag-1 autocorrelation seems to largely be a function of the duration of progress monitoring schedules. That is, positive lag-1 autocorrelation seems to be more of an issue for progress monitoring cases that last longer than 5 months. Before that point, autocorrelation is much more likely to be negative.

Validity of Simulation Studies

Christ and colleagues (2012; 2013) generated progress monitoring cases that lasted a maximum of 20 weeks, but the extant dataset used to generate model parameters for simulations had a range of 3- 34 weeks. A more appropriate method to derive model parameters may have been to fit LMER models to cases that spanned 6-20 weeks to match the data generation model. Considering a large proportion of progress monitoring cases spanned more than 20 weeks within the extant dataset, where most measurement characteristics seemed to stabilize, may have biased parameter estimates for the data generation model. Future simulations should apply a separate LMER model to cases that spanned 6-20 weeks and compare parameter estimates from that model with the parameters originally presented in Table 1.

CBM-R progress monitoring simulation studies did not manipulate missing-ness when generating cases or analyzing the accuracy of decision rules (Christ et al., 2012; 2013). Within the simulation studies, progress monitoring cases were generated with complete data. Based upon the results of this study, this is a potential threat to the external validity of the simulation studies. Future CBM-R progress monitoring simulation studies should manipulate the missing-ness of observations after generating data and measure the effects on the accuracy of decision rules.

The average SEE value within this study seemed to approximate 8-9 WRCM, indicative of a “good” quality dataset. Further, SEE values seemed to stabilize across durations relatively quickly. The model used for data generation in the simulation studies assumed that distributional properties of SEE values were consistent across time (Christ et al.,). The results of this study provide evidence to support that assumption.

The SE*b* values can be interpreted similarly to the root mean square error (RMSE) values in the simulation studies. Within simulation studies RMSE is generally an index of precision, a summary of the absolute average difference between model predicted values and true values. The results of CBM-R progress monitoring simulation studies suggest that RMSE values are largely influenced by duration (Christ et al., 2013). Further, after a certain number of weeks a point of diminishing returns begins to emerge. Based upon the SE*b* values observed in this study, it appears that duration is highly influential on the precision of slope estimates. This is expected given that the computation of SE*b* values factors both SEE and duration into its computation and SEE values seemed to appear relatively stable across time.

Aberrant values were not considered when deriving model parameters for data generation in CBM-R progress monitoring simulation studies. Further, aberrant observations were not simulated within progress monitoring cases. As crude as the operational definitions of aberrant values were in this study, it is clear that a large proportion of progress monitoring cases are prone to have observations that may have disproportionate influence on interpretations of intervention effects.

Growth, or rate of improvement, was assumed to be monotonic and linear within CBM-R progress monitoring simulation studies. Without considering missing observations for short durations, it is difficult to discern whether this is a major cause for concern for the data generation model in CBM-R progress monitoring simulation studies.

Within CBM-R progress monitoring simulation studies, the variability of random effects for true growth did not differ as a function of time. Within multi-level modeling, random effects reflect the magnitude of variability of individual growth estimates or ROI from the group average. It may behoove researchers to conduct follow up simulation studies using multiple data generation models. More specifically, the magnitude of variability for random effects for slope could be greater at shorter durations.

Based upon the results of this study, lag-1 autocorrelation does not appear to be a major threat to the interpretation of CBM-R progress monitoring data. If teachers are supposed to formatively assess student progress after a reasonable length of time, for instance after two or three months, lag-1 autocorrelation is likely to be negative. As a consequence, since data were only generated through 20 weeks within CBM-R progress monitoring simulation studies, the fact that lag-1 autocorrelation was not manipulated does not seem to be a major cause for concern.

Implications for Future Research

Based upon the results of this study, it appears that measurement characteristics that systematically differed the most within the extant dataset was the duration of progress monitoring cases. As a result, future studies that investigate the accuracy of decision rules and visual analysis should manipulate the length of time progress is monitored.

Missing observations were present throughout a substantial portion of cases. If missing observations are considered in light of the duration of progress monitoring schedules, and this extant dataset or other data that is gathered from practice is used in future investigations, it appears that missing data will be inevitable when sampling cases. As a result, at the present time, if investigations of the accuracy of decision rules and visual analysis incorporate actual student data collected in applied practice and duration is manipulated, missing data is likely to occur.

Standard error of the estimate values seemed to stabilize over time. However there was still variability in SEE values when duration was brief. Within the CBM-R research literature different measurement tools and conditions yield different SEE values (Ardoin & Christ, 2009). As a result, it may be prudent to manipulate SEE values in future investigations of time series interpretations. Manipulating SEE values may serve as a proxy for different scenarios where different CBM tools are used under different data collection conditions.

As stated before, aberrant values were poorly defined in this study. More specifically the operational definitions of aberrant values were confounded with variability. A little over 1% of progress monitoring cases could be categorized as not being highly variable and still possessing an extreme value. A more robust operational definition of extreme value should be derived, perhaps drawing from other fields of research, before they are systematically manipulated in investigations of decision rule and visual analysis accuracy with CBM-R.

The magnitude and variability of slope estimates differed across durations. Given that the magnitude of growth is the primary factor evaluated when interpreting CBM-R time series data, it should be systematically manipulated when investigating the accuracy of time series interpretations. The linearity of growth patterns seemed to differ when data collection schedules were short. However, several plausible hypotheses, among them the fact that over fitting at those short durations suggests that at the present time linearity may not be a worthwhile factor to manipulate at the present time.

Finally, autocorrelation, specifically positive lag-1 autocorrelation only seemed to be an issue for extremely long progress monitoring durations. If teachers are expected to make an instructional decision in less than 5 months, it is unlikely that positive lag-1 autocorrelation will distort the interpretation of treatment effects. As a result, as long as investigations into the accuracy of decision rules or visual analysis do not include durations that span upwards of 6 or more months, it is unlikely that autocorrelation is a worthwhile factor to manipulate.

Limitations

Several limitations associated with the current study are worthy of attention. First, the results of this study were completely descriptive. That is, no significance tests were conducted. For instance it is unclear if the difference in the variability of slope estimates at week 3 and 10 are in fact statistically significant.

Second, little was known about the extant dataset analyzed in this study. Student level demographics were not available for each progress monitoring case. As a result, comparisons of duration of progress monitoring schedules SEE values, slope magnitudes, or linearity of growth estimates could not be investigated as a function of student characteristics. Further, data were collected across multiple schools. As a result, observations were not only nested within students, but students were nested within schools. In essence a level of nesting is not being accounted for in any of the analyses.

Related to the second point, student level information regarding the type, intensity, and fidelity of the Tier II interventions they were receiving was not available. This is a larger issue with the CBM-R literature. Little is known how growth estimates, and other pertinent measurement characteristics of CBM-R time series data differ as a function of intervention being delivered and the intensity and fidelity in which that intervention was delivered.

Conclusion

Several measurement characteristics of CBM-R progress monitoring data were explored in this study. Missing-ness, the frequency of aberrant observations, and the severity of autocorrelation within CBM-R progress monitoring data received minimal attention prior to this investigation. Other measurement characteristics such as variability of observations, linearity of growth estimates, and the magnitude of intercept and slope estimates have been investigated before, but have rarely been interpreted as a function of the duration of progress monitoring schedules. Related to the research questions posed at the outset of the study, there was not a typical duration of progress monitoring, instead the distribution of progress monitoring durations was multi-modal. Missing-ness was pervasive and nearly unavoidable, particularly as the duration of progress monitoring increased. Standard error of the estimate values varied across cases, but did not seem to fluctuate as a function of duration. Aberrant values were common, particularly when the duration of progress monitoring was extensive and the SEE values were above average. The linearity of growth patterns seemed to change as a function of time. That is, when students were monitored longer, growth estimates were more likely to be linear. The variability of growth and intercept estimates were high when duration was short and decreased and variability decreased as a function of time. Finally, autocorrelation values did not appear to approach critical values across all durations.

Based upon the results of this study, a key assumption associated with recent CBM-R progress monitoring simulation studies was supported. Distributional properties of residual, or SEE, appeared to be relatively consistent across time. However, the results of this study indicate that additional simulations may be necessary. In particular, future investigations should generate model parameters for data generation based upon durations that more closely reflect the simulated cases that will be analyzed (e.g., 6-20 weeks). In addition, future simulations should model missing-ness and aberrant observations, as both seem to reflect realities of practice.

Summarizing measurement characteristics of CBM-R progress monitoring data is useful, but educators and school psychologists need to know how to interpret the data they collect. That is, to more directly improve formative assessment methods, the results of this study need to inform investigations of the accuracy of visual analysis and CBM-R decision rules. Given the wide array of progress monitoring durations within this extant dataset, the accuracy of visual analysis and CBM-R decision rules need to be investigated as a function of the number of weeks progress is monitored. To that end, investigations should also manipulate measurement characteristics that differ substantially as a function of duration or have been documented as being influential on visual analysis or decision rules. As a result, the duration of progress monitoring cases, the magnitude of observed growth, and the variability of observations (SEE) were manipulated in Study 2.

Chapter 4

Study 2

Students in the same classroom often benefit differently from the same instruction. Educators and researchers promote student success by identifying the unique needs of pupils and delivering interventions to address those needs (Tilly, 2008; Ysseldyke, et al., 2006). Deno (1990) put forward the idea that educators need to act as problem solvers to identify and individualize interventions to address the unique needs of students. To act as problem solvers, educators establish a continuous feedback loop by collecting performance data frequently and evaluating that data regularly to determine the effectiveness of interventions. If a student does not improve the problem solver makes a change to the student's instructional program.

Progress monitoring is broadly defined as the assessment and evaluation of instruction and intervention effects across time (Hixson et al., 2013). Related to the point above, continuous measurement of student progress is not sufficient to improve academic achievement (Mirkin & Deno, 1979; Wesson, 1991). Educators need to use systematic guidelines, often with the help of an experienced practitioner, to evaluate when and if to change instruction (Fuchs & Fuchs, 1986; Fuchs et al., 1989, Stecker et al., 2005).

Curriculum Based Measurement

Curriculum based measurement of oral reading (CBM-R) is often cited as an assessment especially suited to formatively assess instruction across relatively brief periods of time (Deno et al., 1986). In fact, researchers originally developed CBM to evaluate the effectiveness of special education programming for individual students (Deno, 1985; 2003). As a review, when educators use CBM-R, they administer a grade level passage of connected text and calculate the number of words read correctly in one minute (WRCM; Deno, 1985). Educators then administer alternate forms across time, plot the results on time series graphs, and evaluate patterns of observations to make decisions to continue or modify interventions (Deno, 1986). In essence, when an educator collects and evaluates CBM-R progress monitoring data they test how effectively an intervention improves student academic achievement. Further, educators use repeated measures of student performance as an index of that improvement. (Deno, 1985; Deno & Mirkin, 1977).

Researchers and educators use visual analysis (Skiba, et al., 1989) as well as decision rules (Ardoin et al., 2013; Shinn, 1989) to signal the need for instructional modifications when using CBM-R. In theory, interpretative guidelines should limit the likelihood of incorrect decisions (Shapiro, 2011). However, the accuracy of visual analysis and decision rules have received little attention within the context of CBM-R progress monitoring (Ardoin et al., 2013). Further, analyses of large extant datasets (i.e., the results of Study 1) suggest that measurement characteristics that influence the stability and precision of growth estimates systematically differ across students and schedules of progress monitoring. Visual analysis and decision rules are reviewed briefly in the next section

Visual Analysis

Visual analysis remains the predominant method to interpret treatment effects within single case design frameworks (Ximenes, Manolov, Solamas, & Quera, 2009). To visually analyze time series data, a behavior is measured repeatedly in the absence of an intervention (often referred to as a baseline phase). Next, an intervention is introduced and data is again collected across time (often referred to as a treatment phase). Depending on the complexity of the design and the behavior measured, the procedure alternates between baseline and treatment phases (Gast, 2010). As observations are collected, clinicians simultaneously interpret differences in level, trend, and variability of data in the absence and presence of interventions (Kazdin, 1982). Depending on the behavior, a change in level, trend, or variability across phases is indicative of intervention effects, which suggest a functional relationship between the intervention and measurements (Sidman, 1960).

Proponents of visual analysis argue that it: (1) precludes the need for complex statistical analyses, (2) yields accurate interpretations of treatment effects, and by extension (3) results in the retention of clinically significant interventions (Kratochwill & Brody, 1978). Related to points 2 and 3, researchers argued that when using visual analysis to interpret time series data, the likelihood of Type I error is minimal (Franklin, Gorman, Beasley, & Allison, 1997). In this context, Type I error refers to the likelihood that one concludes an ineffective intervention caused a behavior change. Conversely, Type II error refers to an instance where one concludes an effective intervention did not cause a behavior change. Supporters argue that visual analysis protects against Type I errors because behavior change within single case designs is often dramatic (Kazdin, 2011). As a result, visual analysis is thought to result in the retention of interventions that produce not only statistically, but practically significant effects (Baer, 1977; Brossart et al., 2006; Horner, Carr, Halle, McGee, Odom, & Wolery, 2005).

Visual analysis and cbm-r. Despite the often cited advantages of visual analysis, it can result in erroneous decisions (DeProspero & Cohen, 1979; Ottenbacher 1986).

However, recent research suggests that the reliability of visual analysis may be higher than what it was decades ago (Kahng et al., 2010). A rich yet conflicting research literature documents various threats to visual analysis.

Most applications of CBM-R progress monitoring reflect a continuous treatment design (Van Norman, et al., 2013). Within a continuous treatment design, baseline data are not gathered for an extended period of time (often one day) and an intervention is administered continuously for the entire duration of data collection (Glass, Wilson, & Gottman, 1975). Whereas changes in level, trend, and variability across phases is indicative of a treatment effect within multi-phase or reversal single case designs, the rapid increase or decrease of observations within a single phase is the primary unit of analysis within continuous treatment designs. Since multiple phases are not used, practitioners do not make causal claims that an intervention is causing a behavior change (Glass et al.,). As a result, much of the research on threats to visual analysis (e.g., degree of mean and trend shift across phases, presence of baseline trend, and variability of observations across phases) is not directly applicable to CBM-R progress monitoring decisions, which are often made with continuous treatment designs. Rather than analysis and inferences derived across phases of baseline and intervention, it is likely that visual analysis of CBM-R progress monitoring data is within a single treatment phase.

Van Norman and colleagues (2013) observed that trend magnitude influenced the probability of a correct decision among visual analysts who interpreted CBM-R progress monitoring data. The authors of the study used a continuous intervention design where one CBM-R observation was collected per week for 10 weeks. The probability of a correct decision was much higher when a progress monitoring case depicted a substantial weekly rate of improvement (ROI) in WRCM per week (3.00) or no improvement (0 WRCM per week). The probability of a correct decision was lowest when ROI was minimal (0.75 WRCM per week). In addition, the authors found that graphic aides, such as goal lines and trend lines, increased the probability of a correct decision across all trend magnitudes. That study was limited because researchers did not manipulate the duration of data collection or the variability of observations. Another limitation of the study was that the performance of visual analysts was not compared to the accuracy of decision rules.

Decision Rules

When educators evaluate CBM-R progress monitoring data, they must not only determine *if* an intervention is effective, but they must determine *how* effective it is. In other words, the presence or absence of an upward ROI or trend in and of itself is not enough to conclude an intervention is effective for a student. Instead, the magnitude of that trend is the basis for the decision of whether or not an instructional modification should be made. Researchers developed decision rules to simplify interpretations of time series data for educators. Rather than rely upon potentially subjective judgments of trend, decisions to modify instructional programs are sometimes determined with explicit decision rules. Those rules compare a student's ROI or growth to an expected ROI tied to a meaningful long-term goal.

The expected ROI can be graphically depicted as an aim line or goal line. The slope of the goal line quantifies the weekly gain in WRCM the student needs to achieve to reach a long-term goal. Using decision rules, in conjunction with a goal line, theoretically, adds some level of structure and consistency to decisions to modify interventions (Shinn, 2008). Two types of decision rules were developed to evaluate intervention effects: (1) data point decision rules and (2) trend line decision rules (Ardoin et al., 2013).

Data point decision rules were originally used in conjunction with precision teaching (White & Harring, 1980). The primary strength of data point decision rules is their simplicity. When using a data point decision rule the educator only looks at the most recent observations. If the last three observations fall below the goal line, an instructional modification is made. If all of the last three observations fall above the goal line, the slope of the goal line is increased (Fuchs, Fuchs, & Hamlett, 1989). If the most recent three observations are distributed both above and below the goal line, the current instructional program is maintained (Cates, & Ditkowski, 2010). There are variations of the goal line rule, which rely on three, four of five of the most recent consecutive data points.

In contrast, trend line decision rules involve computation of a line of best fit. If the computed slope of that line is less steep than the goal line after a certain number of weeks, the teacher is prompted to make an instructional change (Shinn, 1989). Currently, ordinary least squares (OLS) regression is the preferred method of slope calculation (Ardoin et al., 2013; Christ et al., 2012; Good, Shinn & Stein, 1990; Shinn & Good, 1989).

In their literature review of CBM-R decision rules, Ardoin and colleagues (2013) identified over 100 documents that offered suggestions on how to apply data point and trend line decision rules. Recommendations regarding slope calculation as well as the requisite number of observations to collect before the application of a decision rule varied across documents. The most startling finding was that no empirical investigations evaluated the accuracy of CBM-R decision rules prior to simulation studies (e.g., Christ et al., 2013; reviewed in Chapters 2 and 3).

After the publication of the review, Van Norman and Christ (2014) investigated the accuracy of CBM-R data point and trend line decision rules using an extant progress monitoring dataset. The study was not included in Chapter 2 since it was a conference presentation and thus failed to meet initial inclusion criteria. The authors of the study investigated the accuracy of data point and trend line decision rules using a 1.50 goal line for a data collection schedule where one observation was collected once per week. In addition, progress was monitored concurrently with three different passage-types (AIMSweb, DIBELS Next, and FAST). Ordinary least squares slope estimates from 30 weeks of data collection for each measure served as a proxy for true growth for each student, such that if student growth at 30 weeks was less than 1.50 they were classified as making inadequate progress. Conversely, if the OLS slope was greater than 1.50 they were classified as making adequate progress. Next, three point and trend line decision rules were applied at 4, 5, 6, …20 weeks for each student for each measure. Overall accuracy, sensitivity, and specificity were compared across decision rules, measures, and durations of progress monitoring.

Decision accuracy, sensitivity, and specificity did not substantially differ as a function of passage. Trend line decision rules had higher decision accuracy across all durations of data collection compared to three point decision rules. When using three point decision rules, decision accuracy was below chance until 12 weeks of data collection. When using trend line decision rules, decision accuracy exceeded chance after

6 weeks and .80 after 12 weeks. In addition, data point decision rules displayed high levels of specificity across durations at the expense of sensitivity. In other words, three point decision rules had a high rate of Type I errors. In contrast, minimum standards for sensitivity and specificity (.70; Hintze & Silberglit 2005) were observed after 12 weeks using a trend line decision rule. Trend line decisions seemed to function better than data point rules.

Purpose

The purpose of Study 2 was to evaluate the accuracy of visual analysis and CBM-R decision rules to interpret progress monitoring data. A direct comparison of both evaluative methods will yield empirically based recommendations for the appropriate procedure to interpret CBM-R time series data.

Research questions. Two broad lines of inquiry framed the study. First, what is the accuracy of decisions by visual analysts who evaluate progress monitoring data with and without graphic aides (scatter plot, goal line, trend line) across a variety of conditions? More specifically:

(1) To what degree does trend magnitude influence the accuracy of decisions by visual analysts?

(2) To what degree does the duration (number of weeks) data are collected influence the accuracy of decisions by visual analysts?

(3) To what degree does variability of observations affect the accuracy of decisions by visual analysts?

Whereas the first series of research questions investigated the accuracy of visual analysts, the second set of research questions examined the accuracy of data point and trend line decision rules. More specifically:

(4) How does trend magnitude affect the accuracy of decision rules?

(5) How does the duration (number of weeks) data are collected affect the accuracy of decision rules?

(6) How does the variability of observations affect the accuracy of decision rules?

Methods

Participants

Expert panel. Four professors of Educational Psychology served as expert panel members. Each member of the panel had: (1) earned a PhD in a field related to special education or school psychology, (2) conducted research where visual analysis was the primary method to evaluate treatment effects, (3) applied experience working in schools interpreting time series data, and (4) published peer referred empirical studies and presented at national conferences on topics related to single case design or CBM-R progress monitoring.

Visual analysts. A total of 108 visual analysts completed the study. Originally 113 participants logged on to the website, but 3 subjects were excluded for failing to meet inclusion criteria and 2 subjects were excluded for incorrectly evaluating practice graphs. On average, participants had 4.24 years (*SD* = 4.23) experience using CBM-R. Approximately 80% of the sample was female. The average age of individuals was 29.48 years (*SD* = 7.14). A large proportion of the sample identified as White (90%).

Approximately 3% of the sample identified as Asian. A little less than 1% of the sample identified as African American. The same proportion of participants (~1%) also described themselves as American Indian or Alaskan Native, Hispanic or Latino, and "Other". Approximately 3% of the sample preferred not to disclose their race or ethnicity. Most participants resided in Minnesota (37%) followed by Wisconsin (13%) and Iowa (9%). The remaining visual analysts were spread across a wide array of states including: California, Colorado, Georgia, Illinois, Indiana, Kansas, Louisiana, Massachusetts, Michigan, Mississippi, North Carolina, North Dakota, South Carolina, Washington, as well as two Canadian provinces. A sizable portion (54%) of the sample was school psychology graduate students or interns. Approximately 28% of the sample was practicing school psychologists. Other occupations represented in the study included professors (6%), elementary school teachers (6%), school administrators or RTI coordinators (3%), research associates (2%), and mental health workers (1%).

Materials

One hundred eight fully crossed progress monitoring cases were selected from the extant dataset of 3,078 AIMSweb CBM-R progress monitoring cases (analyzed in Study 1). Time was plotted on the x-axis and each tick mark represented one week of data collection. Words read correct per minute were plotted on the y-axis. Each data point represented one WRCM observation from a single CBM-R administration for that week.

Graphs systematically differed along three dimensions: (1) the magnitude of trend, (2) the duration or number of weeks of progress monitoring, and (3) the level of variability in observations. The levels of each characteristic are detailed in the next section. In addition, depending on the group the visual analyst was randomly assigned to, each graph was displayed one of three ways: (1) as a scatter plot, (2) as a scatter plot with a goal line, or (3) as a scatter plot with a goal line and trend line.

Design

This study followed a 5 (evaluation method) x 3 (trend magnitude) x 3 (duration) x 3 (variability) mixed factorial design. The between subjects factor was evaluation method. The levels of the between subjects factor corresponded to the type of graphic aid used for visual analysis and the type of CBM-R decision rule.

The within subjects factors captured characteristics of the progress monitoring cases, or stimuli. The levels (n=3) of each within subject factor were informed by the results of Study 1. Manipulated factors included: (a) trend magnitude, (b) the duration of progress monitoring cases, and (c) the variability of observations. As a result there were 27 (3 x 3 x 3) unique progress monitoring conditions. Four cases were randomly selected from the extant dataset from Study 1 for each combination of within subject factors, resulting in 108 graphs. The primary outcome was whether responses from visual analysts or the recommendations from decision rules matched responses from an expert panel.

Evaluation method. Five evaluation methods were investigated. The first three methods involved randomly assigning visual analysts to view all progress monitoring graphs one of three ways. One group viewed every progress monitoring case as scatterplot, another group viewed every progress monitoring case as scatter plot with a 1.50 WRCM improvement per week goal line, and the last group viewed each progress monitoring graph with a 1.50 WRCM improvement per week goal line and OLS trend line. The order progress monitoring graphs were presented were randomized for each participant. Decision rules did not rely upon responses from groups of visual analysts. Instead, three point and trend line decision rules were applied to each progress monitoring case by the author.

Trend magnitude. Trend magnitude, or how steeply observations in the progress monitoring graphs rose across time, was set to one of three levels. Generically, these levels can be described as low, medium and high. Trend magnitude was based upon the results of Study 1, and published normative growth rates for second and third grade students (Deno et al., 2001). The average ROI for the low trend group was 0.48 WRCM per week (*range* = 0.04-0.75). This corresponded roughly to the 5th percentile (*range* =1st -9th percentiles). Medium trend magnitude averaged 1.56 WRCM improvement per week (*range* = 1.00-2.00). This approximated the 50th percentile (*range*= 25th – 75th percentiles). Last, high trend magnitude averaged 2.65 WRCM improvement per week (*range* = 2.25 – 3.00). This corresponded roughly to the 95th percentile (*range* = 87th -99th percentiles).

Duration. Three levels of duration were investigated. The levels selected in this study were defined as brief (6 weeks), typical or best practices (12 weeks), and extended (18 weeks). The levels chosen here reflect common durations recommended in the CBM-R progress monitoring literature (Ardoin, et al., 2013) as well as common durations from Study 1.

Variability. Variability of observations was set to one of three levels. In line with Study 1, variability was operationally defined as the magnitude of SEE or residual within a progress monitoring case. The authors of the simulations studies applied qualitative descriptors to four levels of SEE: 5 WRCM = very good, 10 WRCM= good, 15 WRCM

=poor, and 20 WRCM=very poor. While it would be ideal to set four levels of variability in this study, doing so would increase the number of graphs to evaluate considerably (108 versus 144). As a result, three levels of dataset quality were selected: poor, good, and very good. In the current study poor quality datasets had an average SEE value of 14.46 WRCM (*range* = 13.21 – 15.80). Good quality datasets had an average SEE value of 10.03 WRCM (*range* = 9.28-10.94) and very good quality datasets had an average SEE value of 5.18 WRCM (*range* = 3.27-6.60).

Correct responses. Before participants visually analyzed progress monitoring graphs, or decision rules were applied to each case, an expert panel reviewed every progress monitoring graph and reached unanimous agreement whether or not the student was improving at an acceptable rate. As a result, each graph had a criterion. The evaluations of visual analysts and recommendations from decision rules for each case were compared to the decision reached by the expert panel. If the recommendation from a visual analyst agreed with the expert panel for a particular progress monitoring case, a 1 (correct response) was recorded. Conversely, if the recommendation disagreed with the expert panel, the response was coded as 0 (incorrect response). The same process was used to code correct and incorrect responses for each progress monitoring case for decision rules.

Procedure

Expert panel. Each panel member logged onto a website that contained: (a) a brief overview of the study, (b) instructions for responding to each item, and (c) the 108 progress monitoring to evaluate at their own pace. Each graph was presented as a scatterplot, a scatterplot with a goal line, and a scatterplot with a goal line and trend line on a webpage. For each case, experts were asked to determine whether the student was benefitting from their current instructional program at an acceptable rate and the intervention should be continued or the student was not benefitting from their current instructional program at an acceptable rate and another intervention should be implemented. In addition, for each progress monitoring case, panel members were asked to state how confident they were in their evaluation on a scale from 1 (Not at all sure) to 4 (Absolutely sure). The expert panel was unaware of the different ways in which the progress monitoring graphs differed.

Unanimous agreement from the expert panel served as the criterion against which responses from visual analysts and decisions rules were compared. Measuring the correspondence between experts and less experienced raters as well as analytic methods has precedence in industrial organizational psychology (Borman, 1977), school psychology (Riley-Tillman, Chafouleas, Sassu, Chapanese, & Glazer, 2008), and applied behavior analysis (Fisher, Kelley, & Lomas, 2003). The method in which unanimous agreement was reached was adapted from a study that investigated the accuracy of visual analysis (Hagopian, Fisher, Thompson, Owen-DeSchryver, Iawata, & Wacker, 1997). Within the current investigation, experts first evaluated the progress monitoring cases on their own. Next the author identified cases where the experts disagreed. The panel then met to discuss the disagreements. The author asked each member to: (1) make an interpretation of the graph in question, (2) provide the rationale for their decision, and (3) comment on elements of the data that influenced their decision. After that, another vote was taken. If disagreements persisted, the graph was tabled and returned to after other progress monitoring cases were discussed. Initially the panel disagreed on 19 cases

(82.41% agreement). Following the steps outlined above, unanimous agreement was reached for every progress monitoring case after one meeting.

Visual analysts. Visual analysts were recruited several ways. First participation was solicited at a state school psychology conference and a national reading research conference. Second, a research request was distributed on two state school psychology association listservs. Third, alumni from the authors training program were contacted via email and asked to distribute a request to participate to their colleagues and students. Last, colleagues of the author outside of the training program were contacted and asked to distribute the research request to individuals they knew who might be eligible to participate. Subjects who completed the study received a $10 gift card.

Visual analysts logged on to a website to complete the study. After logging on, participants were offered a brief summary of the purpose of the study, the potential risks and benefits of participation, the notification that participation was voluntary and that they may quit at any time. Visual analysts were also informed of how anonymity was protected. After reviewing this information, the portal asked participants if they agreed to participate in the study. Immediately after obtaining informed consent, three screening questions were asked to determine whether or not the individual was eligible to participate. If the participant answered no to any of the following questions the study terminated.

(1) Have you at any point in your career administered and scored a curriculum based measure (CBM)?

(2) During your career, have you collected or oversaw the collection of CBM data across multiple occasions within a semester (i.e., multiple weeks) for an

individual student? In other words have you individually or as a part of a school based group, conducted progress monitoring using CBM? (NOTE: collecting universal screening data alone does not satisfy this requirement.)

(3) During your career have you individually or as part of a school based group looked at a progress monitoring graph and made some kind of decision about student progress (e.g., decided that an intervention was not working)?

Two participants were excluded from the study for answering no to question 3.

After completing the screening questions, visual analysts were introduced to the tasks they would perform. More specifically, participants were told that they would view a series of 108 CBM-R progress monitoring graphs representing students who were receiving supplemental Tier II interventions. Participants were told that they were to assess whether the student was progressing at an acceptable rate or not. If the visual analyst was randomly assigned to a condition with a graphic aid, the graphic aid was defined and example plots with that graphic aid were presented. Next, the visual analysts were informed that for each progress monitoring cases they were to do two things. First, they were to decide whether the student was improving at an acceptable rate and the intervention should be continued, or the student was not improving at an acceptable rate and another intervention should be implemented. Second, they were instructed to indicate how confident they were in their decision on a scale from 1 (Not at all sure) to 4 (Absolutely sure). Two practice cases were then presented. The first graph depicted a clear positive trend and the second case depicted a clear flat trend. If the visual analyst incorrectly stated that the student was not making progress for the first case or incorrectly stated that the student was making adequate progress for the second case, the study terminated. Two participants were excluded from the study for incorrect responses to the practice items.

Decision rules. Decision rules were applied to each progress monitoring case by the author. If the OLS slope for a progress monitoring case was less than 1.50 WRCM per week, the slope of the goal line, a tally was made in a data file indicating the student was making inadequate progress. Conversely if the observed OLS slope was greater than 1.50 WRCM per week a tally was made in the same file indicating that the student was making adequate progress. For the three point decision rule, the author visually inspected each progress monitoring graph. If all of the last three observations fell below the goal line, a tally indicating that the student was making inadequate progress was made. Conversely if any of the final three observations fell above the goal line, a tally was made to indicate the student was making adequate progress.

Analyses

The data-analytic plan was similar to recent visual analysis studies that employed descriptive and inferential analysis (e.g., Christ, Nelson, Van Norman, Chafouleas, & Riley-Tillman, 2013; Nelson et al., 2014; Van Norman et al., 2013). The first series of analyses evaluated the accuracy of decisions from visual analysts against the conclusions drawn from an expert panel. The second series evaluated the accuracy of recommendations from decision rules to the conclusions reached by an expert panel.

Descriptive. The first series of descriptive analyses for visual analysts was conducted at the individual participant level. Each response, as stated before, was coded as correct or incorrect indicating whether the visual analyst agreed or disagreed with the expert panel. First, chi-square fit statistics were computed for each rater collapsed across all characteristics of progress monitoring cases. This analysis was conducted to ensure visual analysts evaluated the 108 progress monitoring cases correctly more often than would be expected by chance. If a participant interpreted the progress monitoring cases correctly at a rate lower than chance (an insignificant chi-square value), the responses from that participant were discarded.

Next, the proportion of correct decisions for each evaluative method was calculated for all possible combinations of trend magnitude, duration, and dataset quality. To adjust for chance agreement, Kappa (Cohen, 1960) was also computed.

Inferential. The first set of inferential analyses was performed to predict how other similar visual analysts would perform with other similar time series graphs. That is, in the first set of analyses the interest was in generalizing across visual analysts and progress monitoring cases. The second series of analyses predicted how decision rules would perform when applied to other similar progress monitoring graphs. The second series of inferential analyses did not aim to generalize across raters, only progress monitoring cases.

The analytic model to predict performance for visual analysts had to account for nesting effects since the same visual analysts evaluated multiple graphs. The analytic model to predict performance for the decision rules did not. As a result, generalized linear mixed modeling (GLMM), an extension of the general linear model, was used predict the average log odds of a correct decision for visual analysts conditioned on the type of graphic aid used as well as different progress monitoring conditions. Logistic regression was used to model the average log odds of a correct response for decision rule conditioned on different progress monitoring conditions. All analyses were conducted with the R computer program (R Development Core Team, 2009)

Results

After initial analyses, the original data analytic plan was updated for two reasons. First, descriptive and inferential results were extremely similar. It seems reasonable to assume that the time series graphs interpreted in this study were sufficiently representative of progress monitoring graphs interpreted in practice. Further, the individuals who participated in this study were sufficiently representative of other individuals who interpret progress monitoring data. As a result, inferential results are discussed for the remainder of the study and descriptive results are presented in Appendix A.

Second, descriptive and inferential analyses became problematic when results were conditioned on trend magnitude. The expert panel, visual analysts, and decision rules were in perfect or near perfect agreement when trend magnitude was high. In such instances descriptive results suggested that agreement was due purely to chance. That is, Kappa was equal to 0. When conducting inferential analyses, particularly for decision rules, standard errors associated with partial slopes for trend magnitude were extremely large. In addition, convergence issues occurred when modeling trend magnitude for the GLMM. The general recommendation when marginal sums or prevalence rates distort omnibus measures of agreement was to report two proportions of specific agreement (Breenan & Prediger, 1981); Positive Agreement (PA) and Negative Agreement (NA). Positive agreement was the probability that for a randomly selected case the result from visual analysis or decision rules agreed with the expert panel that that the student was making inadequate progress. Negative agreement was the probability that for a randomly selected case the result from visual analysis or decision rules agreed with the expert panel that the student was making adequate progress. Trend magnitude was dropped as a predictor from inferential analyses and replaced with a dummy predictor capturing the expert panels evaluation of the progress monitoring case. Modeling the expert panel's evaluation allowed for the estimation of the average log odds of a correct decision when a case depicted inadequate progress (PA) or adequate progress (NA). Inferential results from the original data analytic plan are presented in Appendix B.

The GLMM was used to predict the average log odds of a correct decision for visual analysts conditioned on the type of graphic aid statistically controlling for progress monitoring conditions. Logistic regression was used to predict the average log odds of a correct decision for each decision rule while statistically controlling for progress monitoring conditions. Model fit was computed iteratively for each series of analyses using differences in the Akaike Information Criterion (AIC). Finally, average log odds from final models were converted to probabilities and summarized across evaluation methods and progress monitoring conditions.

Visual Analysts

A null model was fit to serve as a baseline to evaluate the relative improvement in model fit as categorical covariates were added. Within GLMM, a null model, or a model without predictors, is equivalent to a one way random effects analysis of variance:

$$\log\left[\frac{P_{ij}}{1-P_{ij}}\right] = \eta_{ij} = \beta_{0j} + r_{ij} \qquad \text{Level 1}$$

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \qquad \text{Level 2}$$

$Y_{ij} = \gamma_{00} + \mu_{oj} + r_{ij}$ Full (Mixed) Model

Where $\log\left[\frac{P_{ij}}{1-P_{ij}}\right] = \eta_{ij}$ is the log odds that participant *i* correctly interprets progress monitoring case *j* and γ_{00} is the average log odds of a correct interpretation of progress monitoring case *j* across all visual analysts (the fixed effect). Further, μ_{0j} represents the variance in the average log odds of a correct decision from the group level average across visual analysts [(μ_{0j}~N$(0, \tau_{00})$, which is the random effect. Last r_{ij} is residual. The results of the null model (see Table 6) indicated that on average (or when τ_{00}=0), across all types of graphic aids and progress monitoring conditions, the average probability of a correct response was .78 [($e^{1.57}$)/(1+$e^{1.57}$)].

Next, the type of graphic aid was modeled as a categorical covariate (see Table 6). The intercept represented an instance where visual analysts interpreted progress monitoring cases as a scatter plot across all levels of trend magnitude, duration, and dataset quality. On average, the probability of a correct response for the scatter plot condition was .73 as compared to .76 for goal line and .82 for goal line with trend line. Overall, adding graphic aid as a categorical covariate significantly improved model fit relative to the null model ($\chi^2_{(df=4)}$ = 37, p<.05).

The next model included the expert panel's evaluation as a categorical covariate (see Table 6). The intercept represented an instance where visual analysts interpreted progress monitoring cases as a scatter plot across all levels of trend magnitude, duration and dataset quality when a student was making adequate progress (NA). Given those conditions, the average probability of a correct response was .80 compared to .63 when the student was making inadequate progress (PA). Overall, adding the expert panel's evaluation as a categorical covariate significantly improved model fit relative to Model A $(\chi^2_{(df=2)} = 791$, p<.05).

Model C added duration (see Table 6). The intercept represented an instance where visual analysts interpreted progress monitoring cases as a scatter plot, across all levels of trend magnitude, dataset quality, after six weeks of data collection when a student was making adequate progress. Given those conditions, the probability of a correct decision was .77 compared to .80 after twelve weeks and .85 after eighteen weeks of data collection. Overall, adding duration as a categorical covariate significantly improved model fit relative to Model B $(\chi^2_{(df=4)} = 105$, p<.05).

Next, Model D included dataset quality as a categorical covariate. The intercept represented an instance where visual analysts interpreted progress monitoring cases as a scatterplot, across all levels of trend magnitude, dataset quality was good, after six weeks of data collection when a student was making adequate progress. Given those conditions the probability of a correct decision was .77 compared to .70 for poor quality datasets and .85 for very good quality datasets. Overall, adding dataset quality as a categorical covariate significantly improved model fit relative to Model C $(\chi^2_{(df=4)} = 183$, p<.05).

The equation for Model D is presented below:

$$\log\left[\frac{P_{ij}}{1-P_{ij}}\right] = \eta_{ij} =$$

$$\beta_{0j} + \beta_{1j}(Goal\ Line) + \beta_{2j}\,(Goal\ and\ Trend\ Line) +$$

$$\beta_{3j}(Inadequate\ Progress)\ +$$

$$\beta_{4j}(12\ Weeks) + \beta_{5j}\,(18\ Weeks) + \beta_{6j}\,(Poor\ Dataset\ Quality) +$$

$$\beta_{7j}\,(Very\ Good\ Dataset\ Quality) + r_{ij} \qquad \text{Level 1}$$

$\beta_{0j} = \gamma_{00} + \mu_{0j}$

$\beta_{1j} = \gamma_{10} + \mu_{1j}$

$\beta_{2j} = \gamma_{20} + \mu_{2j}$

$\beta_{3j} = \gamma_{30} + \mu_{3j}$

$\beta_{4j} = \gamma_{40} + \mu_{4j}$

$\beta_{5j} = \gamma_{50} + \mu_{5j}$

$\beta_{6j} = \gamma_{60} + \mu_{6j}$

$\beta_{7j} = \gamma_{70} + \mu_{7j}$ Level 2

$\log\left[\frac{P_{ij}}{1-P_{ij}}\right] = \eta_{ij} =$

$\gamma_{00} + \mu_{0j} + \gamma_{10}(Goal\ Line) + \mu_{1j} + \gamma_{20}(Goal\ \ and\ Trend\ Line) + \mu_{2j} +$

$\gamma_{30}(Inadequate\ Progress) + \mu_{3j} + \gamma_{40}(12\ Weeks) + \mu_{4j} + \gamma_{50}(18\ Weeks) +$

$\mu_{5j} + \gamma_{60}(Poor\ Dataset\ Quality) + \mu_{6j} + \gamma_{70}(Very\ Good\ Dataset\ Quality) +$

$\mu_{7j} + r_{ij}$ Full (Mixed) Model

Table 6

Multilevel Analysis Predicting Log Odds of a Correct Decision for Visual Analysts

	Model									
	Null		A		B		C		D	
Predictor	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z
Fixed										
Intercept	1.57 (.05)	34.08*	1.28 (.05)	24.50*	1.80(.07)	16.54*	1.61 (.09)	16.51*	1.60 (.10)	14.71*
Graphic Aid										
GL			.20 (.10)	2.09*	.18 (.09)	2.01*	.16 (.07)	1.96*	.22 (.09)	2.30*
GL & TL			.66 (.08)	8.18*	.75 (.08)	8.71*	.71(.08)	7.95*	.73 (.08)	8.64*
Progress										
Inadequate					-1.12 (.07)	-8.44*	-1.20 (.13)	-9.07*	-1.38 (.13)	-10.24*
Duration										
12 Weeks							.20 (.07)	2.81*	.23 (.07)	3.10*
18 Weeks							.65 (.07)	8.80*	.73 (.08)	9.93*
Dataset Quality										
Poor									-.35 (.06)	-5.70*
Very Good									.63 (.06)	8.59*
Random	*SD*		*SD*		*SD*		*SD*		*SD*	
Intercept	0.40		.20		.60		.76		.81	
Graphic Aid										
GL			.56		.58		.54		.55	
GL & TL			.36		.33		.27		.31	
Progress										
Inadequate					1.26		1.25		1.26	
Duration										
12 Weeks							.34		.36	
18 Weeks							.34		.33	
Dataset Quality										

Poor									.23	
Very Good									.07	
Model Fit	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)
	10733	-	10696	37^* (4)	9905	791^* (2)	9800	105^* (4)	9617	183^* (4)

* p<.05

Note. GL – Goal Line and GL & TL – Goal Line and Trend Line. The intercept in Model D contained the following conditions: Graphic Aid – Scatter Plot, Adequate Progress, Duration – 6 Weeks, Dataset Quality – Good

Decision Rules

Logistic regression was used to predict the average log odds of a correct interpretation for each decision rule while statistically controlling for progress monitoring conditions. A null model was calculated to serve as a baseline to evaluate the relative improvement in model fit as categorical covariates were added. The null model is reproduced below:

$$\log\left[\frac{p}{1-P}\right] = \beta_0 + \varepsilon$$

Where $\log\left[\frac{p}{1-P}\right]$ is the log odds that either decision rule will result in a correct interpretation. β_0 is the intercept estimate and ε is residual. The results of the null model (see Table 7) indicated that on average across both decision rules and all characteristics of progress monitoring cases, the average probability of a correct response was .77 $[(e^{1.58})/(1+e^{1.58})]$.

Next, the type of decision rule was modeled as a categorical covariate (see Table 7). Within Model E the intercept represented the average log odds of a correct decision using a three point decision across all levels of trend magnitude, duration, and dataset quality. On average, the probability of a correct decision for the three point decision rule was .77. The average probability of a correct decision with a trend line was .78, p > .05. Modeling decision rule did not improve model fit relative to the null model ($\chi^2_{(df=1)} = 0$, p>.05).

The next model included the expert panel's evaluation as a categorical covariate (see Table 7). The intercept represented an instance where a three point decision rule was applied across all levels of trend magnitude, duration and dataset quality when a student was making adequate progress (NA). Given those conditions, the average probability of a correct response was .83 compared to .65 when the student was making inadequate progress (PA). Overall, adding the expert panel's evaluation as a categorical covariate significantly improved model fit relative to Model E ($\chi^2_{(df=1)}$ = 6, p<.05).

Model G added duration. The intercept represented an instance where a three point decision rule was applied across all levels of trend magnitude, dataset quality, after six weeks of data collection when a student was making adequate progress (See Table 7). Given those conditions, the probability of a correct decision was .80 compared to .84 after twelve weeks and .86 after eighteen weeks of data collection. Overall, adding duration as a categorical covariate did not improve model fit relative to Model F ($\chi^2_{(df=2)}$ = 2, p<.05).

Next, Model H included dataset quality as a categorical covariate (see Table 7). The intercept represented an instance where a three point decision rule was applied across all levels of trend magnitude, when dataset quality was good, after six weeks of data collection when a student was making adequate progress. Given those conditions the probability of a correct decision was .81 compared to .71 for a poor quality datasets and .91 for a very good quality dataset. Overall, adding dataset quality as a categorical covariate significantly improved model fit relative to Model F ($\chi^2_{(df=4)}$ = 19, p<.05). Model H is reproduced below.

$$\log\left[\frac{p}{1-P}\right] = \beta_0 + \beta_1(Trend\ Line) + \beta_2(Inadequate\ Progress) + \beta_3(12\ Weeks) + \beta_4(18\ Weeks) + \beta_5(Poor\ Quality\ Dataset) + \beta_7(Very\ Good\ Qualtiy\ Dataset) + \varepsilon$$

Table 7

Logistic Regression Analysis Predicting Log Odds of a Correct Response for Decision Rules

	Model									
	Null		E		F		G		H	
Predictor	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z
Intercept	1.58 (.18)	8.73*	1.54 (.25)	6.11*	2.03 (.32)	6.33*	1.76 (.39)	4.51*	1.83 (.51)	3.80*
Rule										
Trend Line			.07 (.36)	.18	.07 (.37)	.19	.07 (.37)	.19	.07 (.39)	.19
Progress										
Inadequate					-1.14 (.37)	-3.08*	-1.20 (.38)	-3.17*	-1.56 (.41)	-3.75*
Duration										
12 Weeks							.36 (.45)	.92	.42 (.46)	.90
18 Weeks							.58 (.46)	1.27	.77 (.48)	1.60
Dataset Quality										
Poor									-.70 (.43)	-1.53
Very Good									1.13 (.56)	2.01*
Model Fit	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)
	200	-	200	0 (1)	194	6 (1)*	196	2 (2)	173	19*(4)

* p <.05

Note. Rule – Decision Rule modeled. The intercept in Model H consisted of the following conditions: Decision Rule –Three Point, Adequate Progress, Duration – 6 Weeks, Dataset Quality – Good

Comparison of Methods

Across all evaluation methods and levels of duration and dataset quality, the average probability of a correct decision was substantially lower when a student was making inadequate progress .67 (PA) compared to when a student was making adequate progress .85 (NA). When a student was making inadequate progress, the average probability of a correct decision was much higher when using a goal line and trend line to visually analyze progress monitoring data (.74, *SD* = .09; see Table 8) relative to other graphic aids and decision rules. In fact, the average probability of a correct decision for other evaluation methods approached chance when a student was making inadequate progress across many progress monitoring conditions. For instance when the quality of dataset was poor after six weeks of data collection, and the student was making inadequate progress, the average probability of a correct decision for a three point decision rule, trend line decision rule, and scatterplot was .42, .43, and .47 respectively.

For the condition described above, the average probability of a correct response for the three point decision rule, trend line decision rule and visual analysis with a scatterplot increased to .71, .72, and .72 respectively when a student made adequate progress. When a student was making adequate progress, the average probability of a correct decision was highly similar across all evaluation methods. For instance, when a student was making adequate progress after 12 weeks with a good quality dataset the average probability of a correct decision for a three point decision rule, trend line decision rule, visual analysis with a scatter plot, visual analysis with an goal line, and visual analysis with a goal line and trend line the average probabilities of a correct decision .85, .86, .81, .83, and .88 respectively.

Table 8

Comparison of the Average Probability of a Correct Decision across Evaluation Methods and Progress Monitoring Conditions

			Evaluation Method				
True	Duration	Dataset		Trend			Goal
Status	(Weeks)	Quality	3 Point	Line	Scatter	Goal	Trend
Adequate Progress (Negative Agreement)	6	P	.71	.72	.72	.76	**.82**
		G	.80	.81	.78	.80	**.86**
		VG	.91	.91	.85	.87	**.92**
	12	P	.77	.78	.76	.79	**.85**
		G	.85	.86	.81	.83	**.88**
		VG	.93	**.94**	.87	.89	.92
	18	P	.85	.86	.82	.85	**.89**
		G	.91	.91	.86	.88	**.92**
		VG	.95	**.96**	.91	.92	.95
M			*.85*	*.86*	*.82*	*.84*	*.89*
SD			*.08*	*.08*	*.06*	*.05*	*.04*
Inadequate Progress (Positive Agreement)	6	P	.42	.43	.47	.52	**.61**
		G	.55	.56	.54	.58	**.68**
		VG	.75	.76	.66	.70	**.77**
	12	P	.50	.51	.52	.56	**.65**
		G	.63	.64	.59	.63	**.71**
		VG	.80	.81	.70	.73	**.82**
	18	P	.63	.64	.61	.65	**.74**
		G	.75	.76	.68	.71	**.79**
		VG	.88	.88	.78	.80	**.89**
M			*.66*	*.67*	*.62*	*.65*	*.74*
SD			*.15*	*.15*	*.10*	*.09*	*.09*
Cases			0%	11%	0%	0%	89%

Note. P- Poor, G- Good, VG- Very Good quality datasets. Bolded values indicate that the evaluation method resulted in the highest probability of a correct decision for that combination of progress monitoring conditions. Cases refers to the percentage of unique combinations a given evaluation method yielded the highest average probability of a correct decision.

Discussion

The purpose of this study was to evaluate the accuracy of two categories of evaluation methods for CBM-R progress monitoring data: visual analysis and decision rules. Inferential analyses were computed to predict the average probability of a correct response when other visual analysts viewed similar time series graphs and decision rules were applied to other similar progress monitoring cases. The effects of duration, and dataset quality on the average probability of a correct decision were explored when students were making inadequate progress (PA) and adequate progress (NA) across all evaluation methods. In addition, the performance of each evaluation method was compared across all progress monitoring conditions.

Visual analysis using a goal line and trend line resulted in higher average probabilities of correct decisions compared to visual analysis with a goal line and visual analysis with a scatter plot. That result was replicated across all progress monitoring conditions. Although it was not expected, the results also indicated that three point decision rules and trend line decisions performed very similar. Overall, visual analysis with a goal line and trend line resulted in the highest average probability of a correct decision when a student was not making adequate progress (.74) and when a student was making adequate progress (.89). Results support the recommendation to use visual analysis of time series graphs with both goal line and trend lines. Further, decisions based on short data collection schedules (less than six weeks) are unlikely to yield accurate interpretations, especially if residuals exceed 10 WRCM and a student is making inadequate progress. Specific patterns observed for each progress monitoring condition are discussed in the next section.

Trend Magnitude

When results were evaluated as a function of trend magnitude, a clear pattern emerged. Visual analysts and CBM-R decision rules were much more accurate and the average probability of a correct decision was near 1.00 when growth was high. Larger magnitudes of growth corresponded with growth rates at the 95th percentile (Deno et al., 2001). However, modeling growth introduced problems with inferential analyses for visual analysts and decision rules. As a result, trend magnitude was dropped and replaced with a covariate capturing whether the student was making inadequate progress or adequate progress as evaluated by the expert panel. The effect of trend magnitude is explored in more depth in Appendices A and B.

Across each visual analytic method and CBM-R decision rule, the probability of a correct decision was much higher when a student was making adequate progress (NA). Further, when a student making inadequate progress (PA), the probability of a correct decision approached chance levels for some evaluation methods when data was collected for six weeks and dataset quality was poor or good. This finding was in direct contrast to a primary tenant of single case design. That is, visual analysis and decision rules were more likely to predict that a student was benefitting from an intervention (Type I error) than to suggest an intervention should be changed. In other words visual analysis and decision rules did not lead to conservative evaluations of treatment effects.

Duration

Across visual analysts, the average probability of a correct decision increased as a function of time. That finding was replicated within the adequate and inadequate growth conditions. The effect of duration was much more pronounced across evaluation methods when a student was making inadequate progress (PA). Further, the incremental improvement in the average probability of a correct decision increased more from 6 to 18 weeks then from 6 to 12 weeks for visual analysts and decision rules.

Variability

Dataset quality, or residual, influenced the accuracy of decision rules particularly when data collection schedules were brief. Both the trend line and data point decision rules had a low probability to yield a correct interpretation when a student was not improving adequately, dataset quality was poor, and data were collected for six weeks. A main effect was observed for dataset quality across all evaluation methods. As dataset quality improved from poor, to good, to very good, average probabilities of a correct decision increased for each evaluation method. This finding emphasizes the importance of high quality data sets with low residual, which are likely a function of good instrumentation and well-standardized administrations and setting across progress monitoring occasions.

Implications for Practice

The findings of this study have several implications for practice. First and foremost, if practitioners only use CBM-R decision rules to guide treatment decisions for individual students as opposed to visual analysis with a goal and trend line, they are at a heightened risk of concluding ineffective interventions are working. Educational professionals need to visually analyze CBM-R progress monitoring data to make a decision for individual students. With increasingly sophisticated computer interfaces, researchers and practitioners can summarize growth estimates for groups of students in a table. If practitioners only evaluate numerical estimates of growth or if they rigidly apply decision rules they are likely to make an incorrect decision.

Second, when employing visual analysis, a goal line and trend line should be used. With the advent of a computer technology, an OLS trend line can be computed in a few keystrokes. Further, almost all CBM-R vendors provide OLS trend lines when producing progress monitoring graphs. Additional in-service training may be required to ensure educators and school psychologists employ trend lines for empirically supported uses (e.g., summarizing current and past performance versus forecasting future performance).

The average probability of a correct decision was strongly related to the duration of data collection, particularly for visual analysis. The more observations visual analysts had to interpret, the higher the average probability of a correct decision. This finding does not challenge the degree to which CBM-R is sensitive to instructional effects across brief periods of time. Instead, the findings of this study suggest that *if* CBM-R is in fact sensitive to changes across brief periods of time, our current evaluative methods for detecting that change are not highly accurate across such short durations. Educators and school psychologists are much more likely to correctly interpret progress monitoring graphs if they collect data for more than six weeks, particularly if the student is showing inadequate progress.

Last, dataset quality or residual influenced the probability of a correct decision across all evaluation methods. Since all students were monitored with the same passage set, differences in SEE values, or dataset quality, are likely the results of differences in standardization and other characteristics of the assessment conditions within the extant dataset. The finding that dataset quality significantly influenced the average probability of a correct decision across all evaluation methods and durations of data collection underscores the importance of collecting data with high fidelity.

Implications for Research

The results of this study also have several implications for research. First, authors of single case design studies that use CBM-R to measure the effects of a single intervention need to be mindful of the graphic aids they are using. That is, before conducting a study using CBM-R, researchers need to identify an expected ROI. They should also supplement progress monitoring graphs with trend lines. Extending beyond a continuous single case design, if a researcher only uses the magnitude of a within phase slope to signal a phase change (e.g., return to baseline), they may be prematurely terminating an intervention and hindering their ability to demonstrate experimental control.

Second, authors who use a continuous intervention design need to be mindful of how long the intervention is delivered. Monitoring progress for less than six weeks with CBM-R is not advisable. Further, researchers need to take steps to reduce residual. They can do this by selecting passage sets with low average levels of SEE and ensuring data are collected in appropriately standardized conditions.

Third, three point decision rules performed much better in this study than in previous simulation studies (Christ, Monaghen, & Balow, 2012). This could be in part because the simulation studies only evaluated magnitudes of trend that approximated the 50th percentile. Christ and Van Norman (2014) used simple derivations from probability theory to suggest that data point decision rules could yield correct decisions at a high rate across relatively brief periods of time when true growth or true ROI substantially differed from the slope of the goal line.

Future research. Future research should systematically manipulate the presence of extreme or aberrant values when evaluating the accuracy of visual analysis and CBM-R decision rules. Outliers or aberrant values were not systematically manipulated in this study. This was in part because of the failure to derive a high quality operational definition in Study 1. However, many of the judgments of the expert panel were influenced by the identification of aberrant values. That is, some expert panel members ignored observations that substantially differed from the array of observations when determining the student was or was not making progress. In some instances, this required the expert panel to ignore the slope of the computed trend line. Although the identification of outliers was confounded with the variability of observations within Study 1, future research needs to improve methods of identifying aberrant values so researchers can investigate their impact on the accuracy of visual analysis and CBM-R decision rules.

Related to the issue of aberrant values, future studies should also investigate the use of other slope calculation methods within the context of visual analysis and CBM-R decision rules. Ordinary least squares regression is highly influenced by aberrant observations at the beginning and end of data series. Non-parametric approaches to slope calculation such as Theil-Sen may be more robust to highly variable data and outliers. In turn, Theil-Sen may improve the accuracy of trend line decision rules and visual analysis relative to OLS based regression.

One data collection schedule was used in this study. A single observation was collected once per week. Although this is the most common CBM-R progress monitoring schedule used in schools, it is unclear if the results of this study would be affected if data were collected multiple times per week or if multiple observations were collected per occasion. Perhaps visual analysts would be overwhelmed with so many observations plotted on a time series graph. Conversely, slope estimates from OLS regression may be more precise if more observations were used to calculate growth. Future studies should manipulate the number of CBM-R observations per week and the number of observations collected per occasion.

Last, future research should investigate the accuracy of decisions and probability of a correct decision when multiple individuals interpret progress monitoring graphs. This study asked individuals to interpret progress monitoring graphs on their own. However, within schools, decisions to modify instructional programming are rarely made independently. Instead, teams of educational professionals meet to discuss individual students and follow a formal problem solving process (Burns, Vanderwood, & Ruby, 2005). As a part of that process, educators collect CBM-R progress monitoring independently and return to the team with the results to discuss potential courses of action. It is unclear whether agreement indices and the average probability of a correct decision would change as a function of the number of people interpreting progress monitoring cases at once.

Limitations

Several limitations are worth noting in the current study. Unanimous agreement among an expert panel was used as a criterion to determine whether a student was or was not making progress. Some standard of true growth, or parallel growth estimates from another measure would have been ideal. However, the use of the expert panel reflected the current state of affairs within educational assessment. That is, there is a paucity of tools with strong psychometric properties that measure instructional effects for individual students across brief periods of time. Further, the extant dataset cases were drawn from did not contain assessment data besides CBM-R observations. In other words, unanimous agreement of the expert panel was the best available criterion for the current study.

Another limitation relates to the inferential analyses used in the study. First, partial slopes for the original series of logistic regression analyses had extremely large standard errors. In addition, only main effects were evaluated in this study. That is, the effect of each independent variable statistically controlling for other predictors was modeled. Interactions were not explored. This is primarily a result of the nature of the independent variables. Since each variable was modeled as a categorical covariate, estimating interactions came at a great cost to statistical power, particularly for the logistic regression analysis. When interactions were modeled within the GLMM, model convergence issues occurred. The same model fitting problems were observed within logistic regression analyses. Future studies that employ more visual analysts and more progress monitoring cases can be conducted so that interactions can be adequately explored.

The sample of visual analysts was drawn from numerous states. However, over half of the visual analysts were school psychology graduate students and interns. Given the limited number of participants, characteristics of visual analysts could not be modeled in the GLMM. It is unclear if the results of this study would be different if only practicing school psychologists or special educators participated. Future research should employ large samples of practicing school psychologists and special educators to determine if graduate students and interns systematically differ in their evaluations of CBM-R progress monitoring graphs.

Last, the manner in which visual analysts answered items may have influenced results. Each visual analyst made a dichotomous decision for each case: the student was improving at an acceptable rate (continue the intervention) or the student was not improving at an acceptable rate (make an instructional change). It could be argued that multiple categories would have been more appropriate. For instance, the student is not improving at an acceptable rate (increase the intensity of the intervention) or the student is improving at an acceptable rate (discontinue the intervention). It is certainly true that when educators interpret progress monitoring data in schools they are capable of making more than two decisions. The dichotomy was chosen in the present study for several reasons. First, increasing the number of response categories would have increased the complexity of analyses considerably. Further it is likely that unanimous agreement would have been much more difficult to attain within the expert panel if multiple response categories were used.

Conclusion

The results of this study affirm what proponents of single case design and visual analysis has been stating for years. Using statistical methods in isolation to guide treatment decisions is not advised (Fisher et al., 1997). Rather, visual analysis should be supplemented with graphic and statistical aids. Indeed, the most accurate evaluative method in this study was visual analysis with an aim line and trend line. However, the findings of this study also conflict with a long-standing belief of single case design. That is, proponents of visual analysis claim that when using the method to interpret time series data, behavior change is dramatic. As a result, visual analysis promotes the retention of only the most clinically significant interventions. The opposite finding was observed within this study. Visual analysts and decision rules had a propensity to over identify students as benefitting from interventions. The pattern was exacerbated when progress monitoring durations were short and the variability of observations was high. As a consequence, when educators use CBM-R progress monitoring data to guide treatment decisions, they are at a high risk of continuing ineffective interventions or inferring that a student is responding to instruction.

The results of this study suggest that educators can take specific steps to minimize the risk of incorrectly interpreting CBM-R progress monitoring data. First, all progress monitoring graphs should be visually analyzed with a goal line and trend line. Decision rules should not be applied automatically. Second, educators should take steps to minimize the variability of observations. Using commercial high quality passage sets and collecting data under highly standardized conditions can minimize the variability of observations and can improve the probability of correct interpretations. Last, practitioners must collect data for longer than six weeks. If educators make such decisions based upon six weeks of data, they are at an especially heightened risk to infer that the student is positively responding to instruction and may incorrectly withhold more effective interventions. In summary, the results of this study suggest that existing CBM-R decision rules and methods of visual analysis can be used to make accurate decisions when data are collected for a sufficient amount of time and educators take steps to minimize the level of variability of observations.

Chapter 5

Integrated Discussion

Deno and colleagues developed CBM around a simple principle. Repeated measures of student performance from grade level materials can be used to index the effectiveness of academic interventions and guide decisions to modify instructional programming (Deno, 1985; 1986). Curriculum based measurement was not developed because of a lack of educational measures. Rather, CBM was developed because most educational measures before it had limited utility in formatively assessing instructional programs (Deno, 2003). Curriculum based measurement was developed as part of a larger data based decision making movement aimed at empowering educators to act as problem solvers (Espin, McMaster & Rose, 2012). The shift helped spark a major change in the science and practice of school psychology. Instead of administering a battery of tests to a student, placing the student in an instructional program, and hoping they would benefit, educators could modify a student's instructional program proactively rather than waiting for them to fail (Deno, 1995).

To be more concrete, progress monitoring is not constrained to CBM. The author of this project feels compelled to disclose that he is a strong advocate for assessing functional skills to guide instructional decisions. The results of this project could be misconstrued as condemning CBM-R progress monitoring practices. However, this project was conducted to shed light on the current evidence for CBM-R evaluation methods. By identifying conditions for which there is poor evidence to use CBM-R as a progress monitoring tool, the author hopes to spur discussion and stimulate research to not only improve CBM-R progress monitoring practices, but formative assessment globally. With those considerations in mind, the results of this project have meaningful implications for advancing CBM-R progress monitoring and formative assessment.

Study 1

In Study 1 seven characteristics were summarized within a large extant CBM-R progress monitoring dataset. More specifically, the variability of observations within cases, the missing-ness of observations within cases, the variability, the frequency of aberrant values, the linearity of growth patterns, the magnitude of intercept and slope estimates, and the magnitude of autocorrelation were computed and summarized across progress monitoring durations.

Several broad implications for school psychologists and educators acting as problem solvers emerged from Study 1. First, researchers and educators need to interpret measurement characteristics of CBM-R progress monitoring data as a function of the duration of data collection. That is, measures of central tendency of several characteristics (precision of growth estimates, frequency of aberrant values, linearity of growth estimates, variability of growth estimates, and severity of autocorrelation) differed across cases that spanned different lengths of time. As a result, omnibus statements related to measurement characteristics of CBM-R progress monitoring data may be misleading.

Relatedly, a few results within Study 1 (e.g., the linearity of growth patterns) conflicted with results of CBM-R progress monitoring studies published elsewhere. Within Study 1 growth was modeled at the individual student level and then summarized across cases. Other investigations of measurement characteristics of CBM-R progress monitoring data often do not summarize growth as it is interpreted in schools. In other words approaches such as multi-level modeling, latent growth modeling, and multiple regression analyses using pooled student data are used to describe CBM-R progress monitoring data. Educators and school psychologists need to remember that the measurement characteristics reported in studies depend in part on the modeling technique used. Consequently, the results of different studies may have different levels of relevance for school based practitioners. The first author chose to fit an OLS regression line to each progress monitoring case and summarized resulting measurement characteristics across cases. This method of growth modeling was selected because it most closely reflected the way data are analyzed in schools. To improve interpretations of CBM-R progress monitoring data, interpretative methods need to be investigated across measurement characteristics likely to be observed in practice.

The results of Study 1 also suggested that the external validity of recent simulation studies (Christ et al., 2012; Christ Monaghen et al., 2013; Christ, Zoploglu et al, 2013) was supported in some aspects and questionable in others. The average level of SEE or variability of WRCM estimates remained relatively constant across progress monitoring durations. That was a key assumption within each of the simulation studies. In addition, lag-1 autocorrelation values did not appear severe enough to undermine the findings of the simulation studies. However, future simulation studies should model missing-ness and the presence of aberrant values. Missing-ness was so pervasive that it appears to be a reality of practice. Aberrant values, although poorly defined in the study, did occur and their potential impact on the accuracy of time series interpretations should be investigated.

While the results of Study 1 are useful in their own right, educators need to know how to interpret the data they collect. Almost a decade ago researchers were called upon to move beyond correlational and descriptive CBM studies (Fuchs, 2004). As a result, progress monitoring cases that differed along three dimensions: the magnitude of trend or ROI, the duration of progress monitoring, and level of variability of observations, were selected and interpreted in Study 2.

Study 2

Ardoin and colleagues (2013) concluded that through 2010, research did not support the use of existing CBM-R decision rules to evaluate instructional programming for individual students across brief periods of time. Relatedly, Van Norman and colleagues (2013) demonstrated that the probability of a correct decision barely exceeded chance levels using visual analysis with CBM-R time series data across certain progress monitoring scenarios. Within Study 2, five methods to interpret CBM-R progress monitoring data were compared across different levels of trend magnitude, durations, and variability. The five evaluation methods included: (1) visual analysis without graphic aids, (2) visual analysis with an expected ROI or goal line, (3) visual analysis with a goal line and trend line, (4) three point decision rule, and (5) a trend line decision rule.

The results of Study 2 suggested that the viability of using CBM-R to measure instructional effects depends in large part on the behavior of educators. That is, educators and school psychologists have a direct influence on the feasibility of CBM-R to guide educational decisions. Educational professionals can take several steps to improve the defensibility of CBM-R as a progress monitoring tool. First, researchers and practitioners should not use decision rules in isolation. Educational professionals that are experienced with CBM-R and visual analysis need to inspect time series graphs to improve the probability of a correct interpretation. Further, practitioners and researchers should visually analyze said time series graphs with goal lines and trend lines. Second, professionals need to take steps to reduce the variability of CBM-R progress monitoring observations. In particular, educators need to make a commitment to use high quality commercial passage sets and consistently collect data in highly standardized conditions. Third, educators need to collect data for longer than six weeks.

The last point has broad implications for progress monitoring and formative assessment. Namely, when using CBM-R or similar measures to monitor student progress for short periods of time, educators are at a heightened risk to conclude ineffective interventions are working. Often, when CBM-R progress monitoring is conducted in schools, a positive response to intervention results in decreased frequency of measurement. The results of Study 2 suggest that it may be necessary to conduct follow up assessment if a student appears to be improving at an acceptable rate after only six weeks. The implication may seem obvious, but is worth stating. Adequate improvement in and of itself is not necessarily a signal that progress monitoring is no longer necessary. Instead, educators should be slightly skeptical of the effectiveness of an instructional program if a student appears to be improving at an acceptable rate after a brief period of time.

Relatedly, the findings of Study 2 converge with the results of Christ et al., 2013. That is, instructional effects need time to substantiate. It should be noted that the criterion for adequate progress in Study 2 (consensus from an expert panel) was drastically different from the criterion for adequate growth in CBM-R progress monitoring simulation studies (simulated true growth; Christ et al., 2012; Christ et al., 2013). The finding that duration was highly influential on the accuracy of time series interpretations across both studies is noteworthy. That is convergent evidence suggests that visual analysis and CBM-R decision rules, in their current state, are not suited to make educational decisions across brief periods of time (less than six weeks).

Conclusion

Although the results of this study suggest there is weak evidence for using CBM-R progress monitoring data to make instructional decisions across brief periods of time, researchers and educators need to remember that progress monitoring is not synonymous with CBM. Efforts to refine and optimize CBM-R instrumentation and data collection procedures are certainly worthwhile. At the same time, progress monitoring and problem solving should not be abandoned if weak evidence continues to emerge for CBM-R progress monitoring practices. Researchers and educators need to remember that CBM is one method of progress monitoring, and alternative methods to measure improvement in academic skills areas are needed. Even if CBM is no longer used, or is drastically different from its current form in 20, 30, or even 40 years, it seems likely that the practice of using instructionally relevant data to guide treatment decisions will result in improved outcomes for students.

References

Albano, A. D., & Rodriguez, M. C. (2012). Statistical equating with measures of oral reading fluency. *Journal of School Psychology*, *50*, 43-59.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Ardoin, S. P., Christ, T. J., Morena, L., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of recommendations and research surrounding curriculum based measurement of oral reading fluency (CBM-R) decision rules. *Journal of School Psychology, 51,* 1-18.

Aron, A., Aron, E., & Coups, E. (2009). *Statistics for Psychology (*4th Ed). Pearson

Baer, D. M. (1977). Perhaps it would be better not to know everything. *Journal of Applied Behavior Analysis, 10,* 167-172.

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

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24,* 339-353.

Breenan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687-699.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, *30*, 531-563.

Burns, M. K., Scholin, S. E., Kosciolek, S., & Livingston, J. (2010). Reliability of decision making frameworks for response to intervention for reading. *Journal of Psychoeducational Assessment. 28*, 102-114.

Burns, M. K., Vanderwood, M. L., & Ruby, S. (2005). Evaluating the readiness of pre-referral intervention teams for use in a problem solving model. *School Psychology Quarterly*, *20*, 89-105.

Cates, G. L., & Kitkowski, B. (March 2010). Ten momments of data-based decision making for RTI schools. Workshop at National Association of School Psychologists: Chicago, IL

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

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

Christ, T. J., & Coolong-Chaffin, M. (2007). Interpretations of curriculum-based measurement outcomes: Standard error and confidence intervals. *School Psychology Forum: Research in Practice*, 1, 75-86.

Christ, T. J., & Hintze, J. M. (2007). Psychometric considerations of reliability when evaluating response to intervention. In S. R. Jimmerson, A. M. Vanderheyden, &

M. K. Burns (Eds.), Handbook of response to intervention (pp. 93−105). New York: Springer.

Christ, T. J., Monaghen, B. D., & Zopluoglu, C. (2012). Diagnostic accuracy of cbm-r data point decision rules. *Unpublished manuscript.*

Christ, T. J., Monaghen, B. D., Zopluoglu, C., Van Norman E. R. (2013). Curriculum-based measurement of oral reading: Evaluation of growth estimates derived with pre-post assessment methods. *Assessment for Effective Intervention, 38,* 139-153.

Christ, T. J., Nelson, P. M., & Van Norman, E. R. (2014). Classical test theory and oral reading fluency. In The Fluency Construct.

Christ, T. J., Nelson, P. M., Van Norman, E. R., Chafouleas, S. M., Riley-Tillman, C. R. (2013). Direct behavior rating: An evaluation of time-series interpretations as consequential validity. *School Psychology Quarterly,*

Christ, T. J., Silberglitt, B., Yeo, S., & Cormier, D. (2010). Curriculum-based measurement of oral reading: An evaluation of growth rates and seasonal effects among students served in general and special Education. *School Psychology Review*, *39,* 447-462.

Christ, T. J., Zopluoglu, C., Long, J. D., & Monaghen, B. D. (2012). Curriculum-based measurement of oral reading: Quality of progress monitoring outcomes. *Exceptional Children, 78,* 356-373.

Christ, T. J., Zopluoglu, C., Monaghen, B. D., & Van Norman, E. R. (2013). Curriculum-based measurement of oral reading: Multi-study evaluation of schedule, duration and dataset quality on progress monitoring outcomes. *Journal of School Psychology, 51,* 19-57.

Christ, T. J., Zopluoglu, C., & Monaghen, B. D. (2012). Diagnostic accuracy of cbm-r trend line decision rules. *Unpublished manuscript.*

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa II: Resolving paradoxes. *Journal of Clinical Epidemiology, 43,* 551-558.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 19,* 37-46.

Colon, E. P., & Kranzler, J. H. (2006). Effect of instructions on curriculum-based measurement of reading. *Journal of Psychoeducational Assessment, 24,* 318-328.

Cronbach, L. J. (1947). Test "reliability": Its meaning and determination *.Psychometrika*, *12*, 1-16.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373-399.

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

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

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

Deno, S. L. (1995). School psychologist as a problem solver. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology III* (pp. 471-484). Washington DC: National Association of School Psychologists.

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 Measurements to Establish Growth Standards for Students with Learning Disabilities. *School Psychology Review*, *30*, 507-524.

Deno, S. L., Marston, D., & Tindal, G. (1986). Direct and frequent curriculum-based measurement: An alternative for educational decision making. *Special Services in the Schools*, *2*(2-3), 5-27.

Derr-Minneci, T. F. (1990). A behavioral evaluation of curriculum-based assessment for reading: Tester, setting, and task demand effects on high- vs. average- vs. low-level readers. *Dissertation Abstracts International, 51*(5-B), 2669.

DeProspero, A. & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 285-296.

Espin, C. A., McMaster, K. L., & Rose, S. (Eds.). (2012). *A measure of success: The influence of curriculum-based measurement on education*. University of Minnesota Press.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43,* 543-549.

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., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1997). Graphical display and visual analysis. In R. D. Frankling, D. B. Allison, & B. S. Gorman

(Eds.), *Design and analysis of single case research* (pp. 119-158). Mahwah, NJ: Lawrence Erbaum.

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research & Practice*, *18*, 172-186.

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. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 33*, 199-208.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989b). Effects of alternative goal structures within curriculum-based measurement. *Exceptional Children, 55*, 429-438.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989c). Monitoring reading growth using student recall: Effects of two teacher feedback systems. *The Journal of Educational Research, 83,* 103-110.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement, using a reading maze task. *Exceptional Children, 58,* 436-450.

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*, 27-27.

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

Gast, D. L. (2010). *Single-subject research methodology in behavioral sciences*. Routledge.

Gersten, R., Compton, D., Connor, C. M., Dimino, J., Santoro, L., Linan-Thompson, S., Tilly, D. (2009). Assisting students struggling with reading: Response to intervention and multi-tier intervention in the primary grades. Institute for Educational Sciences Practice Guide.

Glass, G. V., Wilson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments.* Boulder, CO; Colorado Associate Press.

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

Graney, S. B., Missall, K. N., Martínez, R. S., & Bergstrom, M. (2009). A preliminary investigation of within-year growth patterns in reading and mathematics curriculum-based measures. *Journal of School Psychology*, *47*, 121-142.

Guggenmoos-Holzmann, I. (1993). How reliable are chance-corrected measures of agreement? *Statistics in Medicine, 12,* 2191-2205.

Hadi, A. S., & Simonoff, J. S. (1997). A more robust outlier identified for regression data. *Journal of American Statistics Association, 10,* 281-282.

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

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of r-cbm and high-stakes testing. *School Psychology Review, 34,* 372-386.

Hixson, M. D., Christ, T. J., & Bruni, T. *Best practices in the analysis of progress-monitoring data and decision making*. Chapter under review.

Horner, R. H., Carr, E. G., Halle, J., Mcgee, G. Odom, S., & Wolery, M. (2005). The use of single subject research to identify evidence-based practice in special education. *Exceptional Children, 71,* 165-179.

Huitema, B. E. (1986). Autocorrelation in behavioral research: Wherefore art thou? In A. Poling & R. W. Fuqua (Eds.) *Research methods in applied behavior analysis: Issues and advances* (pp. 187-208). New York: Plenum.

Jenkins, J. R., Graff, J. J., & Miglioretti, D. L. (2009). Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children*, *75*, 151-163.

Jenkins, J., & Terjeson, K. J. (2011). Monitoring reading growth: Goal setting, measurement frequency, and methods of evaluation. *Learning Disabilities Research & Practice*, *26*, 28-35.

Jenkins, J. R., Zumeta, R., Dupree, O., & Johnson, K. (2005). Measuring gains in reading ability with passage reading fluency. *Learning Disabilities Research & Practice, 20,* 245-253.

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (2007). Response to intervention at school: The science and practice of assessment and intervention. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.), *Handbook of response to intervention* (pp. 3-9). New York, NY: Springer.

Kahng, S. W., Chung, K., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 43,* 35-45.

Kane M. T. (2006) Validation In R. L. Breenan (Ed.), *Educational measurement 4th Ed.,* Westport, CT: American Council on Education / Praeger

Kane, M. (2012). Validating score interpretations and uses: Messick lecture, language, testing research colloquium, Cambridge, April 2010. *Language Testing*, *29*, 3-17.

Kane, M. T. (2013). Validating the interpretations and uses of test scores.*Journal of Educational Measurement*, *50*, 1-73.

Kazdin, A. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Kazdin, A., E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.) Oxford: Oxford University Press.

Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification*, *2*, 291-307.

Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using r.* Los Angeles, SAGE Publications, Inc.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23,* 341-351.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Mirkin, P. K. & Deno, S. L. (1979). *Formative evaluation in the classroom: An approach to improving instruction.* (Research Report No. 10).

Nelson, P. M., Van Norman, E. R., & Christ, T. J. (2014). The effect of a brief experimental training on visual analysis with outliers. *Manuscript submitted for publication.*

Nese, J. F., Biancarosa, G., Anderson, D., Lai, C. F., Alonzo, J., & Tindal, G. (2012). Within-year oral reading fluency with CBM: a comparison of models. *Reading and Writing*, *25*, 887-915.

Nese, J. F., Biancarosa, G., Cummings, K., Kennedy, P., Alonzo, J., & Tindal, G. (2013). In search of average growth: Describing within-year oral reading fluency growth across Grades 1–8. *Journal of school psychology*, *51*, 625-642.

Ottenbacher, K. (1986). Reliability and accuracy of visually analyzing graphed data from single-subject designs. *American Journal of Occupational Therapy, 40,* 464-469.

Parker, R. I., Vannest, K. J., Davis, J. L., & Clemens, N. H. (2012). Defensible progress monitoring data for medium-and high-stakes decisions. *The Journal of Special Education*, *46*, 141-151.

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

R Development Core Team (2009). R: A lanague and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www. R-project.org

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, *47*, 427-469.

Riley-Tillman, T. C., & Burns, M. K. (2009). Single case design for measuring response to educational intervention. *New York: Guilford*.

Scruggs, T. E., Mastropieri, M. A., Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education, 8,* 24-33.

Shapiro, E. S. (2008). Best practices in setting progress monitoring goals for academic skill improvement. *Best practices in school psychology IV* (pp. 141-158). Bethesda, MD: National Association of School Psychologists.

Shapiro, E. S. (2011). *Academic skills problems: Direct assessment and intervention*. Guilford Press.

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

Shinn, M. R. (2008). Best practices in using curriculum based measurement and its use in the problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in* *school psychology IV* (pp. 243-262). Bethesda, MD; National Association of School Psychologists.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. (1992). Curriculum-based reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21,* 458-478.

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

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.

Silberglitt, B., & Hintze, J. M. (2007). How Much Growth Can We Expect? A conditional analysis of r cbm growth rates by level of performance. *Exceptional Children*, *74*, 71-84.

Skiba, R., Deno, S., Marston, D., & Casey, A. (1989). Influence of trend estimation and subject familiarity on practitioners' judgments of intervention effectiveness. *The Journal of Special Education*, *22*(4), 433-446.

Skiba, R. J., Deno, S. L., Marston, D., & Wesson, C. (1986). Characteristics of time-series data collected through curriculum-based reading measurement. *Assessment for Effective Intervention*, *12*(1), 3-15.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review*, *30*, 407-419.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using Curriculum‑Based Measurement to Improve Student Achievement: Review of Research. *Psychology in the Schools*, *42*, 795-819.

Ticha, R. (2008). *The effects of pairing curriculum-based measurement with a structured approach that includes providing instructional alternatives on teacher decision making and student achievement*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Tindal, G., Deno, S. L., & Ysseldyke, J. E. (1983). Visual analysis of time series data: Factors of influence and level of reliability. Technical Report 112. Institute for Research on Learning Disabilities. Minneapolis, MN.

Van Norman, E. R., & Christ, T. J. (2014). *CBM-R decision rules: We changed our minds again.* Paper presented at the National Association of School Psychologists Annual Conference. Washington D.C.

Van Norman, E. R., Nelson, P. M. Shin, J. & Christ, T. J. (2013). An evaluation of the effects of graphic aids in improving decision accuracy in a continuous treatment design. *Journal of Behavioral Education, 22,* 283-301.

vanDerHeyden, A. M., Witt, J. C., & Barnett, D. W. (2005). The emergence and possible futures of response to intervention. *Journal of Psychoeducaitonal Assessment, 23,* 339-361.

Vannest, K. J., Parker, R. I., Davis, J. L., Soares, D. A., & Smith, S. L. (2012). The theil-sen slope for high-stakes decisions from progress monitoring. *Behavioral Disorders*, *37*, 271-280.

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.

Wesson, C. L. (1991). Curriculum-based measurement and two models of follow up consultation. *Exceptional Children, 57,* 246-256.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching*. CE Merrill Publishing Company.

Ysseldyke, J., Burns, M., Dawson, P., Kelley, B., Morrison, D., Ortiz, S., et al. (2006). School psychology: A blueprint for training and practice III. Bethesda, MD: National Association of School Psychologists.

Ximenes, V. M., Manolov, R., Solana, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. *The Spanish Journal of Psychology, 12,* 823-832.

Appendix A

Descriptive Analyses

Data from visual analysts were screened. One insignificant chi-square value was observed. This suggests that the visual analyst answered items correctly at a rate less than chance. An insignificant chi-square value indicated low fidelity of task completion so that. That participant's data were removed (<1% of the data).

Visual Analysts

Proportion correct values and Kappa coefficients were computed for all unique combinations of progress monitoring conditions and evaluation methods. After initial analyses, it became clear that marginal sums within contingency tables distorted measures of agreement. That is, in certain scenarios either the expert panel, all visual analysts within a group, or the results of a decision rule rated every case the same (e.g., all cases were making adequate progress). As a result, some marginal sums equaled zero within contingency tables and omnibus measures of association such as Kappa suggested that agreement was due purely to chance. Similar issues with omnibus measures of association have been documented elsewhere (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990). Further some authors challenged the degree to which Kappa truly corrects for chance agreement, particularly because it is based on hypothetical probabilities (Guggenmoos-Holzmann, 1993).

The general recommendation when marginal sums or prevalence rates distort omnibus measures of agreement was to report two proportions of specific agreement (Breenan & Prediger, 1981); Positive Agreement (PA) and Negative Agreement (NA). Positive agreement was the probability that for a randomly selected case the result from visual analysis or decision rules agreed with the expert panel that that the student was making inadequate progress. Negative agreement was the probability that for a randomly selected case the result from visual analysis or decision rules agreed with the expert panel that the student was making adequate progress.

Across all progress monitoring conditions, the proportion of correct decisions for all visual analysts was .82 (PA =.73 , NA=.86). Across all durations and levels of dataset quality, the agreement for low trend magnitude was.74 (PA=.83, NA=.52); medium trend magnitude was.79 (PA=.55, NA=.87), and high trend magnitude was .92 (PA=0, NA=.96).

The proportion of correct decisions for visual analysts at six weeks was .79 (PA=.68, NA=.74), twelve weeks was .82 (PA=.70, NA=.88), and eighteen weeks was .85 (PA=.80, NA=.88). Positive agreement increased more than either overall agreement or NA.

A similar main effect was observed for dataset quality. The proportion of correct decisions for poor quality datasets was .78 (PA=.63, NA=.84) good quality datasets was .82 (PA=.71, NA=.87) very good quality datasets was.86 (PA=.83, NA=.89) for very good quality datasets.

The proportion of correct decisions for visual analysts in the scatter plot condition was .78 (PA = .69, NA= .83), the goal line condition was .82 (PA= .70, NA=.86), and the goal line and trend line condition was .87 (PA=.81, NA=.90). While the proportion of correct decisions increased as the number of graphic aids increased, the degree of improvement in PA was more pronounced. That is, across characteristics of progress monitoring conditions visual analysts made highly accurate evaluations in cases with inadequate growth; however, PA was further improved in conditions with goal and trend lines.

Table 9 delineates the results further for visual analysts. Within the table the proportion of correct decisions, PA, and NA are presented for each visual analyst group, conditioned on duration, dataset quality, and trend magnitude. Across all levels of duration and dataset quality, visual analysts were most accurate when trend magnitude was high. As a result, when dataset quality improved a sharper increase in the proportion of correct decisions more for low and medium trend magnitudes. Further, when dataset quality was very good, the accuracy of visual analysts with goal and trend lines improved considerably as duration increased relative other to other visual analysts.

Decision Rules

Across all progress monitoring conditions the proportion of correct decisions for both decision rules was .83 (PA=.74, .87). Across progress monitoring conditions, the proportion of correct decisions was similar for low (.72, PA=.82, NA=.41) and medium (.76, PA=.48, NA=.85) trend magnitudes. Positive agreement for decision rules improved when trend magnitude was low versus when it was medium. When trend magnitude was high, decision accuracy was high across all other progress monitoring conditions (1.00, PA=0, NA=1.00).

The proportion of correct decisions for each decision rule increased from .79 (PA = .65, .85) at six weeks to .85 (PA = .76, NA = .89) at twelve weeks. Although the overall proportion of correct decisions did not increase from 12 weeks to 18 weeks (.85), PA values did improve (.76 to .80).

For very poor quality datasets, the proportion of correct decisions was .74 (PA = .51, NA=.82). That value improved to .83 (PA = .73, NA=.88) for good quality datasets and to .92 (PA = .90, NA=.93) for very good quality datasets.

The overall proportion of correct decisions were highly similar for the three point decision rule .82 (PA = .67, NA= .88) and the trend line decision rule .83 (PA =.79, NA=.86). Over and above proportion of correct decisions, across all progress monitoring conditions, the trend line rule demonstrated higher levels of PA than the three point decision rule. Table 10 delineates the results for each decision rule conditioned on the duration, quality of datasets, and trend magnitude. When trend magnitude was high, both decision rules yielded perfect agreement. As a result the remainder of the discussion focuses on low and medium trend magnitudes. The three point decision rule showed an inconsistent pattern of results as a function of the duration of progress monitoring. However, the performance of trend line decision rules seemed to show consistent improvement as the duration of progress monitoring cases increased from six weeks .81 (PA = .71; NA = .82) to twelve weeks .81 (PA= 74; NA = .84) to eighteen weeks .92 (PA = .90; NA=.93).

Table 9

Agreement Indices for Visual Analysts

			Overall Agreement (Positive Agreement, Negative Agreement)		
			Trend Magnitude		
Evaluation Method	Duration (Weeks)	Dataset Quality	Low	Medium	High
Scatter Plot	6	P	.45 (.27, .56)	.74 (.59, .80)	.72 (0, .84)
n = 36		G	.84 (.88, .72)	.58 (0, .73)	.70 (0, .82)
.78 (.69, .83)		VG	.87 (.93, 0)	.77 (.77, .77)	.89 (0, .94)
		All	.72 (.79, .57)	.70 (.56, .77)	.77 (0, .87)
	12	P	.64 (.71, .51)	.58 (0, .73)	.83 (0, .91)
		G	.74 (.73, .74)	.76 (0, .86)	.98 (0, .99)
		VG	.92 (.96, 0)	.88 (.73, .92)	.97 (0, .99)
		All	.76 (.83, .62)	.74 (.29, .84)	.93 (0, .96)
	18	P	.71 (.83, 0)	.85 (.66, .90)	.94 (0, .97)
		G	.71 (.83, 0)	.61 (.39, .72)	.99 (0, 1)
		VG	.67 (.80, 0)	.76 (.26, .85)	1 (0, 1)
		All	.70 (.82, 0)	.74 (.44, .83)	.98 (0, .99)
Goal Line	6	P	.69 (.79, .91)	.88 (.24, .81)	.84 (0, .91)
n = 35		G	.80 (.85, .71)	.90 (0, .95)	.72 (0, .84)
.81 (.70, .86)		VG	.53 (.69, 0)	.88 (.87, .89)	.90 (0, .95)
		All	.67 (.70, .64)	.88 (.78, .92)	.82 (0, .90)
	12	P	.53 (.55, .50)	.65 (0, .79)	.77 (0, .87)
		G	.78 (.75, .81)	.77 (0, .87)	.94 (0, .97)
		VG	.86 (.93, 0)	.81 (.56, .88)	.96 (0, .98)
		All	.72 (.78, .63)	.75 (.24, .85)	.89 (0, .94)
	18	P	.76 (.87, 0)	.93 (.86, .95)	.95 (0, .98)
		G	.81 (.89, 0)	.72 (.64, .78)	.97 (0, .98)
		VG	.72 (.84, 0)	.74 (.21, .84)	.98 (0, .99)
		All	.76 (.87, 0)	.80 (.62, .86)	.97 (0, .98)
Goal & Trend	6	P	.53 (.38, .62)	.88 (.77, .92)	.96 (0, .98)
n = 36		G	.76 (.84, .55)	.91 (0, .95)	.94 (0, .97)
.87 (.81, .90)		VG	.78 (.88, 0)	.85 (.82, .87)	.99 (0, .99)
		All	.69 (.77, .53)	.88 (.75, .92)	.96 (0, .98)
	12	P	.71 (.79, .52)	.78 (0, .88)	.97 (0, .99)
		G	.63 (.67, .58)	.94 (0, .97)	.99 (0, 1)
		VG	.97 (.99, 0)	.88 (.71, .92)	.99 (0, .99)
		All	.77 (.85, .54)	.87 (.43, .92)	.98 (0, .99)
	18	P	.89 (.94, 0)	.83 (.57, .90)	.99 (0, 1)
		G	.92 (.96, 0)	.72 (.62, .78)	1 (0, 1)
		VG	.89 (.94, 0)	.85 (.59, .91)	1 (0, 1)
		All	.90 (.95, 0)	.80 (.60, .87)	1 (0, 1)

Note. P- Poor, G- Good, VG- Very Good quality datasets respectively.

Table 10

Agreement Indices for Decision Rules

			Proportion of Overall Agreement (Positive Agreement, Negative Agreement)		
			Trend Magnitude		
Decision Rule	Duration (Weeks)	Dataset Quality	Low	Medium	High
3 Point	6	P	.75 (0, .86)	.75 (0, .86)	1 (0,1)
.82					
(.67, .88)		G	.50 (.50, .50)	1 (0, 1)	1 (0,1)
		VG	.25 (.40, 0)	1 (1, 1)	1 (0,1)
		All	.50 (.40, .57)	.92 (.80, .95)	1 (0,1)
	12	P	.50 (.50, .50)	.75 (0, .86)	1 (0,1)
		G	.75 (.67, .80)	1 (0, 1)	1 (0,1)
		VG	1 (1,0)	1 (1, 1)	1 (0,1)
		All	.75 (.80, .67)	.92 (.67, .95)	1 (0,1)
	18	P	.25 (.40, 0)	.75 (0, .86)	1 (0,1)
		G	1 (1,0)	.50 (0, .67)	1 (0,1)
		VG	.75 (.86, 0)	.75 (0, .86)	1 (0,1)
		All	.67 (.8, 0)	.67 (0, .80)	1 (0,1)
Trend Line	6	P	.25 (.40, 0)	.50 (0, .67)	1 (0, 1)
.83					
(.79, .86)		G	.75 (.86, 0)	.50 (0, .67)	1 (0, 1)
		VG	1, (1,0)	1 (1, 1)	1 (0, 1)
		All	.67 (.80, 0)	.67 (.5, .75)	1 (0, 1)
	12	P	.75 (.86, 0)	.50 (0, .67)	1 (0, 1)
		G	.50 (1, 0)	.75 (0, .86)	1 (0, 1)
		VG	1 (1, 0)	.75 (.67, .80)	1 (0, 1)
		All	.75 (.86, 0)	.67 (.33, .78)	1 (0, 1)
	18	P	1 (1,0)	.50 (0, .67)	1 (0, 1)
		G	1 (1, 0)	.75 (.67, .80)	1 (0, 1)
		VG	1 (1, 0)	1 (1, 1)	1 (0, 1)
		All	1 (1, 0)	.75 (.57, .82)	1 (0, 1)

Note. P-Poor, G-Good, and VG-Very Good quality datasets respectively.

Appendix B

Original Inferential Analyses

The GLMM was used to predict the average log odds of a correct decision for visual analysts conditioned on the type of graphic aid statistically controlling for characteristics of the progress monitoring graphs. Logistic

regression was used to predict the average log odds of a correct decision for each decision rule while statistically controlling for different characteristics of the progress monitoring graphs. Model fit was computed iteratively for each series of analyses using differences in the Akaike Information Criterion (AIC). Finally, average log odds from final models were converted to probabilities and summarized across evaluative methods and characteristics of progress monitoring cases.

Visual Analysts

A null model was fit to serve as a baseline to evaluate the relative improvement as categorical covariates were added. Within GLMM, a null model, or a model without predictors, is equivalent to a one way random effects analysis of variance:

$$\log\left[\frac{P_{ij}}{1-P_{ij}}\right] = \eta_{ij} = \beta_{0j} + r_{ij} \qquad \text{Level 1}$$

$$\beta_{0j} = \gamma_{00} + \mu_{0j} \qquad \text{Level 2}$$

$$Y_{ij} = \gamma_{00} + \mu_{oj} + r_{ij} \qquad \text{Full (Mixed) Model}$$

Where $\log\left[\frac{P_{ij}}{1-P_{ij}}\right] = \eta_{ij}$ is the log odds that participant *i* correctly interprets progress monitoring case *j* and γ_{00} is the average log odds of a correct interpretation of progress monitoring case *j* across all visual analysts (the fixed effect). Further, μ_{0j} represents the variance in the average log odds of a correct decision from the group level average across visual analysts [(μ_{0j}~N$(0, \tau_{00})$, which is the random effect. Last r_{ij} is residual. The results of the null model (see Table 11) indicated that on average (or when τ_{00}=0), across all types of graphic aids and characteristics of the progress monitoring cases, the average probability of a correct response was .78 [($e^{1.57}$)/(1+$e^{1.57}$)].

Next, the type of graphic aid was modeled as a categorical covariate (see Table 11). The intercept model represented a scenario where visual analysts interpreted progress monitoring cases as a scatter plot across all levels of trend magnitude, duration, and dataset quality. On average, the probability of a correct response for the condition of scatter plot was .73 as compared to .76 for goal line and .82 for goal line with trend line. Overall, adding graphic aid as a categorical covariate significantly improved model fit relative to the null model ($\chi^2_{(df=2)} = 37$, p<.05).

Model J added trend magnitude as a categorical covariate (see Table 11). The average probability of a correct response differed depending on whether slope magnitude was low, .65, medium, .71, or high, .88, for scatterplot interpretations. The same pattern was observed when cases were interpreted with a goal line with .68, .73, and .89; and with both goal and trend line with .77, .81, and .93 for low, medium and high trend magnitudes respectively. All parameters were statistically significant (p<. 05). Overall, adding trend magnitude as a categorical covariate significantly improved model fit relative to Model I ($\chi^2_{(df=2)} = 587$, p<.05).

Model K added the duration of data collection as a categorical covariate (see Table 11). When trend magnitude was held constant at medium, as the duration of progress monitoring increased from 6 weeks to 12 weeks to 18 weeks, the average probability of correctly interpreting a progress monitoring case as a scatter plot improved from .68 to .72 to .75 respectively. Overall, adding duration as a categorical covariate significantly improved model fit relative to Model J ($\chi^2_{(df=2)}$ = 66, p<.05).

Model L added the quality of datasets or residual as a categorical covariate (see Table 11). When trend magnitude and duration were held constant at medium and 6 weeks respectively, the average probability of a correct decision was .62 when dataset quality was poor, was .68 when dataset quality was good and was .74 when dataset quality was very good. Overall, adding dataset quality as a categorical covariate significantly improved model fit relative to Model K ($\chi^2_{(df=2)}$ = 86, p<.05).

Decision Rules

Logistic regression was used to predict the average log odds of a correct interpretation for each decision rule while statistically controlling for characteristics of progress monitoring cases. A null model was calculated to serve as a baseline to contextualize the relative improvement in model fit as categorical covariates were added. The null model is reproduced below:

$$\log\left[\frac{p}{1-P}\right] = \beta_0 + \varepsilon$$

Where $\log\left[\frac{p}{1-P}\right]$ is the log odds that either decision rule will result in a correct interpretation. β_0 is the intercept estimate and ε is residual. The results of the null model (see Table 12) indicated that on average across both decision rules and all characteristics of progress monitoring cases, the average probability of a correct response was .77 [($e^{1.58}$)/(1+$e^{1.58}$)].

Next, the type of decision rule was modeled as a categorical covariate (see Table 12). Within Model M the intercept represented the average log odds of a correct decision when applying a three point decision rule across all levels of trend magnitude, duration, and dataset quality. On average, the probability of a correct decision for the three point decision rule was .77. The average probability of a correct decision for the trend line decision rule was .78, p > .05. The decision rule did not improve model fit ($\chi^2_{(df=1)}$ = 0, p>.05).

Model N included trend magnitude as a categorical covariate (see Table 12). The average probability of a correct when slope magnitude was low was .67, medium was .71 or high was .99 when interpreting progress monitoring cases with a three point decision rule. Adding trend magnitude as a categorical covariate significantly improved model fit relative to Model M($\chi^2_{(df=2)}$ = 18, p<.05).

Model O added duration as a categorical covariate (see Table 12). When trend magnitude was held constant at medium, and progress monitoring cases were interpreted with a three point decision rule, the average probability of a correct decision improved was .66 after six weeks, was.73 after twelve weeks, and .74 after 18 weeks. Overall, adding duration as a categorical covariate did not significantly improved model fit relative to Model F ($\chi^2_{(df=2)}$ = 2, p>.05). Adding duration as a categorical covariate did significantly improve model fit relative to the Null Model ($\chi^2_{(df=6)}$ = 26, p<.05).

Model P added the quality of datasets or the magnitude of residual variance as a categorical covariate (see Table 12). When trend magnitude and duration were held constant at medium and 6 weeks respectively, the average probability of a correct decision when using a 3 point decision rule increased was .54 for poor quality datasets .67 for good quality datasets and was .80 for very good quality datasets. Adding dataset quality as a categorical covariate significantly improved model fit relative to Model O ($\chi^2_{(df=2)}$ = 6, p<.05).

Table 11

Original Multilevel Analysis Predicting Log Odds of a Correct Decision for Visual Analysts

	Model									
	Null		I		J		K		L	
Predictor	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z
Fixed										
Intercept	1.57 (.05)	34.08*	1.28 (.05)	24.50*	1.13 (.07)	16.54*	.96 (.09)	11.51*	.95 (.10)	9.88*
Graphic Aid										
GL			.20 (.10)	2.09*	.17 (.09)	2.01*	.16 (.09)	1.89*	.26 (.09)	2.87*
GL & TL			.66 (.08)	8.18*	.72 (.08)	8.71*	.65 (.08)	7.95$^-$	.64 (.08)	7.73*
Trend Magnitude										
Low					-.33 (.07)	-4.61*	-.33 (.08)	-4.46*	-.33 (.07)	-4.39*
High					1.38 (.10)	13.42*	1.38 (.10)	13.53*	1.41 (.10)	13.65*
Duration										
12 Weeks							.24 (.06)	3.48*	.24 (.07)	3.54*
18 Weeks							.47 (.08)	5.94*	.48 (.08)	6.02*
Dataset Quality										
Poor									-.30 (.06)	-4.91*
Very Good									.39 (.07)	5.20*
Random	*SD*		*SD*		*SD*		*SD*		*SD*	
Intercept	0.40		.20		.35		.50		.58	
Graphic Aid										
GL			.56		.24		.38		.43	
GL & TL			.36		.02		.23		.27	
Slope										

Magnitude										
Low					.47		.49		.50	
High					.67		.65		.66	
Duration										
12 Weeks							.31		.33	
18 Weeks							.48		.48	
Dataset Quality										
Poor									.05	
Very Good									.34	
Model Fit	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)
	10733	-	10696	37^* (2)	10109	587^* (2)	10043	66^* (2)	9957	86^* (2)

* p<.05

Note. GL –Goal Line and GL & TL – Goal Line and Trend Line. The intercept in Model D contained the following conditions: Graphic Aid – Scatter Plot, Slope Magnitude - Medium, Duration – 6 Weeks, Dataset Quality – Good

Table 12

Original Logistic Regression Analysis Predicting Log Odds of a Correct Response for Decision Rules

	Model									
	Null		M		N		O		P	
Predictor	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z
Intercept	1.58 (.18)	8.73*	1.54 (.25)	6.11*	1.13 (.33)	3.34*	.86 (.41)	2.0*	.89 (.51)	1.75
Rule										
Trend Line			.07 (.36)	.18	.07 (.38)	.19	.07 (.39)	.19	.08 (.40)	.20
Trend Magnitude										
Low					-.22 (.38)	-.57	-.22 (.38)	-.57	-24 (.40)	.40
High					18.39 (1267)	.02	18.39 (1267)	.02	18.42 (1223)	.02
Duration										
12 Weeks							.43 (.46)	.92	.46 (.48)	.95
18 Weeks							.43 (.46)	.92	.46 (.48)	.95
Dataset Quality										
Poor									-.66 (.44)	-1.53
Very Good									.86 (.55)	1.55
Model Fit	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)
	200	-	200	0 (1)	172	28 (2)*	174	- -	168	6*(2)

* p <.05

Note. Rule – Decision Rule modeled. The intercept in Modell P consisted of the following conditions: Decision Rule –Three Point, Slope Magnitude - Medium, Duration – 6 Weeks, Dataset Quality – Good

Table 12

Original Logistic Regression Analysis Predicting Log Odds of a Correct Response for Decision Rules

	Model									
	Null		M		N		O		P	
Predictor	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z	β (SE)	z
Intercept	1.58 (.18)	8.73*	1.54 (.25)	6.11*	1.13 (.33)	3.34*	.86 (.41)	2.0*	.89 (.51)	1.75
Rule										
Trend Line			.07 (.36)	.18	.07 (.38)	.19	.07 (.39)	.19	.08 (.40)	.20
Trend Magnitude										
Low					-.22 (.38)	-.57	-.22 (.38)	-.57	-24 (.40)	.40
High					18.39 (1267)	.02	18.39 (1267)	.02	18.42 (1223)	.02
Duration										
12 Weeks							.43 (.46)	.92	.46 (.48)	.95
18 Weeks							.43 (.46)	.92	.46 (.48)	.95
Dataset Quality										
Poor									-.66 (.44)	-1.53
Very Good									.86 (.55)	1.55
Model Fit	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)	AIC	χ^2 (df)
	200	-	200	0 (1)	172	28 (2)*	174	- -	168	6*(2)

* p <.05

Note. Rule – Decision Rule modeled. The intercept in Modell P consisted of the following conditions: Decision Rule –Three Point, Slope Magnitude - Medium, Duration – 6 Weeks, Dataset Quality – Good

Comparison Across Evaluation Methods

Regression coefficients for all possible combinations of evaluation methods and progress monitoring conditions within Models L and P were added to derive the average log odds for a correct decision. Those average log odds were then transformed to probabilities. Table 13 presents the average probability of a correct decision

On average, visual analysis with a goal line and trend line, resulted in the highest probability of a correct decision (.83, *SD* =.07). The trend line decision rule generated similar results .80, *SD* = .16). Both decision rules were highly accurate when trend magnitude was high (Mode= .99). However, when trend magnitude was low or medium, visual analysts using a goal line and trend line resulted in the highest average probability of a correct decision across all conditions of duration and dataset quality (.78, *SD* = .05).

Table 13

Average Probabilities of a Correct Decision for each Evaluation Method

			Evaluative Method				
Modeled Characteristic			Decision Rules		Visual Analysis		
Trend Magnitude	Duration (Weeks)	Dataset Quality	3 Point	Trend Line	Scatter	Goal	Goal Trend
Low	6	P	.50	.51	.56	.61	.68
Medium			.54	.56	.62	.67	.73
High			.99	.99	.83	.86	.89
Low		G	.62	.64	.62	.66	.73
Medium			.67	.68	.68	.72	.78
High			.99	.99	.86	.88	.91
Low		VG	.76	.78	.69	.73	.79
Medium			.80	.81	.74	.78	.82
High			.99	.99	.89	.91	.93
Low	12	P	.59	.60	.61	.65	.72
Medium			.63	.65	.67	.71	.77
High			.99	.99	.86	.88	.91
Low		G	.70	.72	.66	.71	.76
Medium			.74	.75	.72	.76	.81
High			.99	.99	.88	.90	.93
Low		VG	.82	.83	.73	.76	.84
Medium			.85	.86	.77	.81	.87
High			.99	.99	.91	.93	.94
Low	18	P	.59	.60	.65	.70	.75
Medium			.63	.65	.71	.75	.80
High			.99	.99	.88	.90	.92
Low		G	.70	.72	.70	.74	.79
Medium			.74	.75	.75	.79	.83
High			.99	.99	.90	.92	.94
Low		VG	.82	.83	.76	.80	.84
Medium			.85	.86	.80	.83	.87
High			.99	.99	.93	.94	.95
M			.79	.80	.75	.79	.83
SD			.17	.16	.11	.10	.07
Conditions			33%	33%	0%	0%	66%
Excluding High Trend Magnitude							
M			.69	.71	.69	.73	.78
SD			.11	.11	.06	.06	.05
Conditions			0%	0%	0%	0%	100%

Note. Conditions refers to the percentage of unique progress monitoring conditions a given evaluative method yielded the highest probability of a correct decision. Totals do not add to 100% because of ties.