Journal of Psychoeducational Assessment
2018, Vol. 36(1) 55–73

DOI: 10.1177/0734282917712174
journals.sagepub.com/home/jpa
SAGE

Curriculum-Based Measurement of Reading: An Evaluation of Frequentist and Bayesian Methods to Model Progress Monitoring Data

Theodore J. Christ[1] and Christopher David Desjardins[1]

Abstract

Curriculum-Based Measurement of Oral Reading (CBM-R) is often used to monitor student progress and guide educational decisions. Ordinary least squares regression (OLSR) is the most widely used method to estimate the slope, or rate of improvement (ROI), even though published research demonstrates OLSR's lack of validity and reliability, and imprecision of ROI estimates, especially after brief duration of monitoring (6-10 weeks). This study illustrates and examines the use of Bayesian methods to estimate ROI. Conditions included four progress monitoring durations (6, 8, 10, and 30 weeks), two schedules of data collection (weekly, biweekly), and two ROI growth distributions that broadly corresponded with ROIs for general and special education populations. A Bayesian approach with alternate prior distributions for the ROIs is presented and explored. Results demonstrate that Bayesian estimates of ROI were more precise than OLSR with comparable reliabilities, and Bayesian estimates were consistently within the plausible range of ROIs in contrast to OLSR, which often provided unrealistic estimates. Results also showcase the influence the priors had estimated ROIs and the potential dangers of prior distribution misspecification.

Keywords

progress monitor, curriculum-based measurement, response to intervention (RtI), multitiered system of supports (MTSS), assessment of interventions/outcomes, education assessment, curriculum-based assessment, bayesian, assessment, measurement, diagnosis

Idiographic evaluation of a student's response to intervention (RtI) is a contemporary approach to guide instructional and diagnostic decisions. RtI informs the diagnosis of learning disabilities and decisions regarding access to prevention, remediation, and special education services (Individuals With Disabilities Education Improvement Act, 2004). This definition of RtI is specific to the collection and use of data before, during, and after an instructional program is implemented to evaluate its effect and the student response. This approach to educational decisions was conceptualized and popularized by the work of Deno (Deno, 1986, 2003; L. S. Fuchs & Deno, 1991) and others (D. Fuchs, Fuchs, McMaster, & Al Otaiba, 2003; Gresham, 2001).

[1]University of Minnesota, Minneapolis, MN, USA

Corresponding Author:
Theodore J. Christ, University of Minnesota, 250 Education Sciences Building, 56 East River Road, Minneapolis, MN 55455, USA.
Email: tchrist@umn.edu

It emerged, in part, in response to the lack of theoretical and empirical support for diagnostically prescriptive instruction (Deno, 1990, 1995). Along with the concept of RtI, Deno (Deno, 1986, 1995, 2003; Deno, Marston, & Tindal, 1985) proposed and promoted the development of curriculum-based measurement (CBM) to facilitate the collection and use of time-series data for RtI.

CBM was developed to index the level and rate of academic performance in the basic skill areas of reading, mathematics, written expression, and spelling (Deno, 1985). A standardized set of procedures were developed to administer and score CBM. From the inception of CBM in the early 1980s until about 2000, the assessment materials were directly sampled from the curriculum materials used in the classroom (L. S. Fuchs & Deno, 1991). It was later established that the inconsistencies of curriculum samples contributed excessive variability to student scores across time, administrations, and alternate forms (Hintze & Christ, 2004; Hintze, Christ, & Keller, 2002). New methods have emerged to select (Christ & Ardoin, 2009; Poncy, Skinner, & Axtell, 2005) and arrange the stimuli (Christ & Vining, 2006) that comprise the 20-to-30 alternate forms, which need to be of substantially similar qualities to ensure consistent measurement across occasions.

The results of a comprehensive literature review indicate that the most commonly published recommendation was to collect one CBM score per week for 6 to 10 weeks (Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013). Those data are then compared with a goal line, which defines the expected rate of improvement (ROI; Ardoin et al., 2013). That goal line was used in conjunction with either a data-point or trend-line procedure.

The data-point procedure indicates that the program effects are sufficient if the most recent data points are distributed either above-*and*-below the goal line or all above the goal line. The program effects are insufficient if the data points are all below the goal line. Historically, there was very little evidence for the technical adequacy of the data-point rule (Ardoin et al., 2013). A subsequent study indicated that the probability of a correct decision using the data-point rule approximated chance, and there was a bias in the decisions derived over the first many weeks of monitoring (Van Norman & Christ, 2016). The bias increased the chances that an ineffective instructional program with low ROI would continue erroneously.

The trend-line procedure indicates the program effects are sufficient if the estimated ROI approximates or exceeds the ROI of the goal line. If not then it indicates that the effects are insufficient. Although there was very little evidence for the technical adequacy of the trend-line rule (Ardoin et al., 2013), it was established that trend estimates were more accurate if statistical methods were used to estimate the ROI as compared with visual analytic analysis or statistical methods (Good & Shinn, 1990; Shinn, Good, & Stein, 1989). Those findings helped shift the emphasis from visual analysis to statistical analysis, specifically employing ordinary least squares regression (OLSR). Later, published findings cast doubt on the precision of the estimated ROIs (Christ, 2006; Hintze & Christ, 2004). Published findings indicate that the residual—or variability of student performance around the trend line—was often too large in magnitude to yield precise estimates of the ROI. From this point forward, the focus is on CBM of Oral Reading (CBM-R), which is the most widely used and researched among the CBM procedures (Wayman, Wallace, Wiley, Ticha, & Espin, 2007).

CBM-R

CBM-R is an individually administered assessment. The examiner reads standardized directions to instruct a student to read aloud from the top of the page. The directions indicate that the students should try to read each word and if they come to a word they do not know then it will be told to them (after 3 s). The examiner times the 1-min reading and marks any incorrectly read words, and then calculates the words read correct per minute (WRCM). There is good evidence

for the reliability of scores and modest evidence for construct and criterion validity (Wayman et al., 2007). There is also modest evidence that student achievement improves under some conditions when teachers collect and use the data to inform instruction (Stecker & Fuchs, 2000; Stecker, Fuchs, & Fuchs, 2005). Notwithstanding, there are challenges for the use of CBM-R to evaluate and inform instructional programs for individual students. One substantial challenge is that it remains very difficult to discern the true trend of student achievement—because there is substantial variability in student scores across time.

Researchers observed that the variation of CBM-R scores around the OLSR line was 11, 12, and 15 WRCM (Dynamic Indicators of Basic Early Literacy Skills (DIBELS); University of Oregon Center on Teaching and Learning, 2017) when they used each of three different probe sets (Ardoin & Christ, 2009) . That observation was generally consistent with previously published findings (Good & Shinn, 1990; Hintze & Christ, 2004), which indicate that the variation of student performances is likely to fluctuate within the range of 5 to 20 WRCM (Christ, 2006). Researchers consistently concluded that the variability in student performance across time was problematic in that it hindered the estimation, interpretation, and use of CBM-R to evaluate RtI.

Published studies indicate that CBM-R scores are highly sensitive to the qualities of the probes (Ardoin & Christ, 2009; Hintze & Christ, 2004), conditions during the administration (Derr & Shapiro, 1989), and disposition of the student. At the time of this study, the evidence for the interpretation and use of CBM-R indicated that 2 or more months of weekly data were necessary to estimate the ROI with OLSR. Although OLSR was the preferred and recommended procedure to evaluate RtI, the precision of those estimates were often insufficient and required extended duration of data collection (e.g., Christ, 2006).

Simulation Studies

A series of published studies reported the methods and findings from simulation studies that explored various methods to estimate ROI from CBM progress monitoring data. The first in the series of studies examined the reliability and validity of trend estimates derived from weekly monitoring (Christ, Zopluoglu, Long, & Monaghen, 2012). The initial results indicated that 14 weeks CBM-R data wer needed to derive valid and reliable estimates of growth usign OLSR. A similar study indicated that pre–post CBM-R scores might yield trend estimates of only slightly less qualities, such that pre-post data collection rather than weekly data collection might be recommended in some scenerios (Christ, Monaghen, Zopluoglu, & Van Norman, 2013).

Finally, a comprehensive paper simulated data across numerous possible conditions to derive estimates of growth with OLSR for broad range of possible data collection conditions (Christ, Zopluoglu, Monaghen, & Van Norman, 2013). The *durations* of collection spanned 2 to 20 weeks. The *schedule* spanned the collection of 1 to 6 CBM-R scores each week, such that some schedules established 3 scores in a single day and others established 1 score per day on each day of the week. Finally, the conditions of *data quality* for the residuals was 5, 10, 15, and 20 WRCM. Those simulated results were analyzed to estimate the reliability, validity, standard error (precision), and diagnostic accuracy of ROIs estimated with OLSR. Results indicate that durations of 10-to-14 or more weeks of data were required in all scenerios. That included a 6-week duration of daily data collection for which there was insufficient technical adequacy for ROIs estimated with OLSR. The results indicated that the OLSR estiamtes of ROI were substantially influenced by the duration of data collection and the quality of the data (i.e., residual). The results associated with the 6-week duration of daily data collection were replicated in field-based study. OLSR-based estimates of ROI are very likely insufficient even with a good quality set of data that was collected every day for 6 weeks field-based study (Thornblad & Christ, 2014). Across studies, results consistently indicate that OLSR estiamtes of ROIs require higher quality data and a longer

duration of data collection. Those are often unrealistic conditions within practice, which requires short durations of data collection to make quick decisions. Moreover, it is rare that CBM-R or other similar measurement procedures result in very high quality data with small residuals. Thus, research and development is necessary to isolate alternate methods to yeild reliable and valid estimates of ROI with moderate quality data collected over brief durations.

Purpose

The purpose of this research was to introduce the concept of Bayesian methods and compare them with OLSR as a competing method to model CBM-R time-series data for individual students. Specifically, it was to illustrate and evaluate the use of Bayesian method across a variety of common conditions for progress monitoring, which spanned various durations (6, 8, and 10 weeks), schedules (weekly and biweekly), and growth rates (low and high/typical growth ROIs). A series of Bayesian models were evaluated with a range of priors to evaluate their influence on estimates of ROI.

Method

Data Generating Mechanism

Data for the simulation were generated from a normal linear model. A normal linear model was selected as OLSR is the standard method for estimating ROIs, which is consistent with the historical (Good & Shinn, 1990) and contemporary practice (Christ, 2006; Christ & Ardoin, 2009; Christ et al., 2012; Christ, Zopluoglu, et al., 2013; Hintze & Christ, 2004) . Given that WRCM is inherently a nonnegative integer, a Poisson or negative binomial model could have been used as the data generating mechanism and subsequently a loglinear regression model could have been fit to the data. Because we were interested in comparing the performance of OLSR with Bayesian methods and not loglinear regression, we decided to use normal-based methods for both data generation and model fitting to give OLSR the best circumstances for performance.

Specifically, we generated data from the following model:

$$Y_{ijkl} \sim N\left(\mu_{ijk}, \sigma_l^2\right), \tag{1}$$

where Y_{ijkl} is the number of WRCM for the jth observation ($j = 1,\ldots,30$) for student i ($i = 1,\ldots,2,000$) with the kth ($k = 1$ or 2) level of student growth (high or low) and the lth ($l = 1$ or 2) condition of data quality (good or poor). In other words, Y_{ijkl} represented the observed WRCM for student i at time j who had either high ($k = 1$) or low ($k = 2$) growth and whose data were of good ($l = 1$) or poor quality ($l = 2$). The levels of student growth and data quality are described below.

The true WRCM, μ_{ijk}, for student i at time j with growth k was a linear function of

$$\mu_{ijk} = \beta_0 + \beta_{1ik} x_j, \tag{2}$$

where x_j represented the week of monitoring, which ran from 1 to 30 weeks. The observed WRCM for this student was

$$Y_{ijkl} = \mu_{ijk} + \varepsilon_{ijkl}, \tag{3}$$

where $\varepsilon_{ijkl} \sim N(0, \sigma_l^2)$. This means that the observed scores were conditionally independent, both within and across students, given the parameters in Equation 2.

Table 1. Conditions of Data Collection.

Condition	Schedule	Duration (weeks)	Number of observations	Observation weeks
1	Biweekly	6	3	1, 3, 5
2	Biweekly	8	4	1, 3, 5, 7
3	Biweekly	10	5	1, 3, 5, 7, 9
4	Biweekly	30	15	1, 3, 5, . . . , 25, 27, 29
5	Weekly	6	6	1, 2, 3, 4, 5, 6
6	Weekly	8	8	1, 2, 3, . . . , 6, 7, 8
7	Weekly	10	10	1, 2, 3, …, 8, 9, 10
8	Weekly	30	30	1, 2, 3, . . . , 28, 29, 30

Parameter Values for the Data Generating Mechanism

In Equation 2, β_0, the intercept, that is, a student's true score at the start of progress monitoring, was fixed to 40 WRCM. This value was selected based on the estimated WRCM at the start of progress monitoring for students in Grades 2 and 3 presented in Christ, Zopluoglu, et al. (2013).

Conditions of differential student growth and quality of the data were considered in this simulation. The student growth (ROI), β_{1ik}, was determined based on whether a student had high (i.e., general population) or low (i.e., special education population) growth. The distribution for high growth was $\beta_{1i1} \sim N(1.4, 0.4^2)$, and the distribution for low growth was $\beta_{1i1} \sim N(.8, 0.4^2)$. Each student had a true ROI that was a realization from one of these two distributions depending on whether they were from the general or special education population (Deno, Fuchs, Marston, & Shin, 2001; L. S. Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993).

The quality of the data set was determined by the magnitude of the residual standard deviation. The quality of the residual standard deviation corresponded to either good, $\varepsilon_{ijk1} \sim N(0, 10^2)$, or poor, $\varepsilon_{ijk2} \sim N(0, 15^2)$. These values correspond to the values reported in several empirical studies (Ardoin & Christ, 2009; Christ, 2006; Christ & Ardoin, 2009; Francis et al., 2008; Good & Shinn, 1990; Hintze & Christ, 2004; Jenkins, Zumeta, Dupree & Johnson, 2005; Jenkins, Graff & Miglioretti, 2009; Shinn et al., 1989) and were the basis for a large simulation study on CBM-R (Christ, Zopluoglu, et al., 2013).

Independent Variables

In addition to student growth and data quality, the duration and the schedule of progressing monitoring and the statistical framework were examined as independent variables.

Duration and schedule. The duration and scheduling of progress monitoring were varied in this study. The duration of progress monitoring was set to 6, 8, 10, or 30 weeks. The scheduling of progress monitoring was either weekly or biweekly. Therefore, there were a total of eight data collection conditions (four durations times two schedules) used for progress monitoring. These conditions are shown in Table 1.

As indicated in the "Data Generating Mechanism" section, 30 observations were generated for each student. Depending on the condition of duration and schedule, only a subset of observations from these original 30 observations were used. For example, the use of all 30 observations for a student would correspond to a weekly schedule with a duration of 30 weeks (Condition 8 in Table 1). If the duration was 6 weeks, then only the first six observations of the complete 30 observations could be used (i.e., the first 6 weeks of the 30 weeks; Conditions 1 and 5 in Table 1). If the monitoring was weekly, then all six observations, observations from Weeks 1, 2, . . . , 5, 6, were used (Condition

5 in Table 1), and if the monitoring was biweekly, then observations from Weeks 1, 3, and 5 were used (Condition 1 in Table 1). This process of subsetting was repeated for the other combinations of duration and schedule.

Statistical Framework

The frequentist and Bayesian approaches were compared in this simulation study.

Frequentist approach. By a frequentist approach, we are referring to the traditional framework used to fit statistical models and conduct hypothesis testing. It is termed frequentist because probability is defined as the frequency of observing an event if a study or experiment was repeated over an essentially infinite number of occasions. This frequency of an event interpretation of probability yields the cumbersome definitions of p values, and confidence intervals that are learned, and struggled with, in introductory statistics courses. Parameters, for example, the ROI, from a frequentist perspective are considered fixed, and all inferences about these parameters are made on their sampling distributions. In a frequentist framework, only the observed data are included in the statistical models. This is in contrast to the Bayesian approach (presented next).

The frequentist model considered in this study was OLSR. Specifically, we fit the following simple linear regression model to the data:

$$Y_{ijkl} = \beta_0 + \beta_{1ik} x_j + \varepsilon_{ijkl}. \tag{4}$$

Because we generated data from the normal linear model with conditionally independent observed scores, the OLSR model in Equation 4 is the same model as the data generating model presented in Equation 3.

Bayesian approach. In a Bayesian approach, statistical conclusions about our unknown parameters, θ, are made in terms of a subjective, personal belief, view of probability conditional on the observed data (Gelman, Carlin, Stern, & Rubin, 2004). These subjective beliefs are formally incorporated into statistical distributions about our unknown parameters or effects (known as prior distributions).

A prior distribution represents our best guess at what the distribution of a parameter or effect would be prior to collecting any data for our study. They are usually developed based on published literature or expert opinion, although in a hierarchical linear model they may be partially developed based on the observed data (known as an empirical prior). If there is a plethora of information about the plausible values of a parameter, we may use a more informative prior with a small variance giving a lot of weight to where we believe the parameter might be, whereas if we are unsure what a parameter could be, we might specify a wide range of possible values for that parameter with relatively equal, low probabilities across the entire range. The former is typically referred to as an informative prior, while the latter would be categorized as a noninformative prior. We discourage this distinction as all prior distributions are informative to varying degrees. The major benefit of less informative priors is that they may converge with the frequentist models and allow a researcher to minimize the influence of a prior distribution, but they do not always converge and what is noninformative on one metric may not be noninformative when transformed (Carlin & Louis, 2009).

In a Bayesian framework, the prior distribution is then multiplied by a likelihood function. A likelihood function corresponds to the probability of the data given all the plausible values for our parameters. The likelihood function represents a traditional statistical model. For example, in this article, our likelihood function corresponds to the OLSR model presented in Equation 4. In Equation 4, there are three different parameters that must be estimated, and therefore, we need to

specify prior distributions for each of these. Prior distributions must be specified for the intercept, β_0, the slope (i.e., ROI), β_1, and the residual variance, σ^2 (though for mathematical and computational convenience, Bayesians typically invert this variance and model the precision, τ).

The resulting distribution from the product of the prior distribution, $p(\theta)$, and the likelihood function, $p(Y_{ijkl}|\theta)$, is known as the posterior distribution, $p(\theta|Y_{ijkl})$, and it is this distribution, the posterior distribution, that all Bayesian inferences are made from. Formally, we arrive at the posterior distribution through Bayes' theorem. Bayes' theorem specifies the relationship between two conditional probabilities and allows us to invert the conditional probabilities. Equation 5 shows Bayes' theorem.

$$p\left(\theta|Y_{ijkl}\right) = \frac{p\left(\theta\right)p(Y_{ijkl}|\theta)}{p\left(Y_{ijkl}\right)}. \tag{5}$$

This states that our posterior distribution, which contains the probabilities of our parameters conditional on the data, is equal to the prior distribution of our parameters multiplied by the probability of the data given our parameters (i.e., the likelihood function), divided by the distribution of our data. The denominator is an integrating constant that allows our posterior distribution to be a true probability distribution (i.e., all probabilities sum or integrate to 1). Because the denominator does not have any parameters, and it is a constant, it is usually omitted and Bayes' theorem is reexpressed as

$$p\left(\theta|Y_{ijkl}\right) \propto p\left(\theta\right)p\left(Y_{ijkl}|\theta\right). \tag{6}$$

Through Equations 5 and 6, we are able to derive the probabilities that are of substantive interest to researchers, that is, the probability of our parameters given the data. The posterior distribution is estimated using Markov chain Monte Carlo (MCMC; Carlin & Louis, 2009; Gelman et al., 2004). Prior to the development and use of MCMC, Bayesian priors were specified in order that they could be solved numerically limiting the range of potential forms of prior distributions. The prior distributions were selected to be conjugate to the posterior distribution, where conjugate means that the posterior distribution is of the same form as the prior distribution (e.g., if a normal distribution was specified for a prior, then the posterior distribution would be normal). However, MCMC frees us from needing to use conjugate priors and allows us to specify prior distributions that may be either parametric (e.g., normal, uniform, or gamma distributions) or nonparametric in nature, which may more realistically represent our view of the parameters. MCMC works by iteratively estimating values for each of our parameters based on the prior distributions and likelihood function, and saves each of these estimates which become the posterior distribution.

Once we have estimated the posterior distribution via MCMC, probabilistic statements about the parameters follow without the need for a null hypothesis statistical testing framework. For example, it is possible to directly calculate the probability of an ROI being greater than 1.1 or 1.4, rather than calculating the probability of observing an ROI or greater given that the null hypothesis is true (i.e., the frequentist *p* value). We can also calculate the probability of the ROI being between 1.1 and 1.4. The Bayesian equivalent to a 95% confidence interval, known as a 95% credible interval, can be directly interpreted as there being a 95% probability that the parameter or effect is between the lower and upper limits of the interval.

Readers who are interested in a more complete introduction to Bayesian statistics and MCMC are directed to the texts of Gelman et al. (2004), Carlin and Louis (2009), and Kruschke (2015), where the former two texts require knowledge of statistical theory and calculus and the latter does not. In addition, Zyphur and Oswald (2015) have written a highly readable introduction to Bayesian statistics that is especially relevant for the testing community.

Table 2. Prior Distributions for β_1.

Prior	*M*	Variance	Precision
High growth, low variance	1.4	0.04	25
High growth, intermediate variance	1.4	0.16	6.25
High growth, high variance	1.4	0.64	1.5625
Low growth, low variance	1.1	0.04	25
Low growth, intermediate variance	1.1	0.16	6.25
Low growth, high variance	1.1	0.64	1.5625

Note. Precision is the inverse of variance, and precision was the parameter used for the prior distributions.

For our simulation, prior distributions must be specified for β_0, β_1, and σ^2 as indicated above. For β_0 and σ^2, we specified priors with large variances over a wide range of plausible values. Specifically, the prior distributions for β_0 and for σ^2 were uniform distributions from 0 to 100 or more compactly, $\beta_0 \sim U(0,100)$ and $\sigma^2 \sim U(0,100)$. The mean and variance of these prior distributions were 50 and 833.33, respectively. Because these were uniform distributions, they gave equal weight to every value between 0 and 100. Because β_1, the ROI term, which, represents expected weekly growth in WRCM, is the focal point in CBM-R intervention, we investigated six different normal prior distributions. The prior distributions are presented in Table 2. These priors differ in their assumed expected growth and variance around this value. All six of these priors on β_1 would be considered highly informative, and the priors with the lowest variance were the most informative.

Dependent Variables

Three dependent variables were selected for this simulation study. They were reliability, mean signed difference (MSD), and root mean square error (RMSE). The squared correlation of the true ROI, β_{1i}, and the estimated ROI, $\widehat{\beta}_{1i}$, is an index of reliability (i.e., the proportion of the variability in the true ROIs that is explained by the estimated ROIs). MSD is a measure of bias, the difference between the expected value of the ROI and the true ROI, and represents systematic deviations from the true ROI. MSD was defined as

$$\mathrm{MSD}\left(\widehat{\beta}_{1i}\right) = \sum_{i=1}^{n} \frac{\widehat{\beta}_{1i} - \beta_{1i}}{n}. \tag{7}$$

Finally, RMSE is a measure of variability between the observed and the true ROI, and was defined as

$$\mathrm{RMSE}\left(\beta_{1i}\right) = \sqrt{E\left(\widehat{\beta}_{1i} - \beta_{1i}\right)^2}. \tag{8}$$

Results

The results are presented first descriptively and then by each individual outcome measure—reliability, MSD, and RMSE. Within a given section, the Bayesian priors are compared with one another and with OLSR. Given the large number of conditions and large number of replicates of each condition, graphical data analysis was the principal form of analysis.

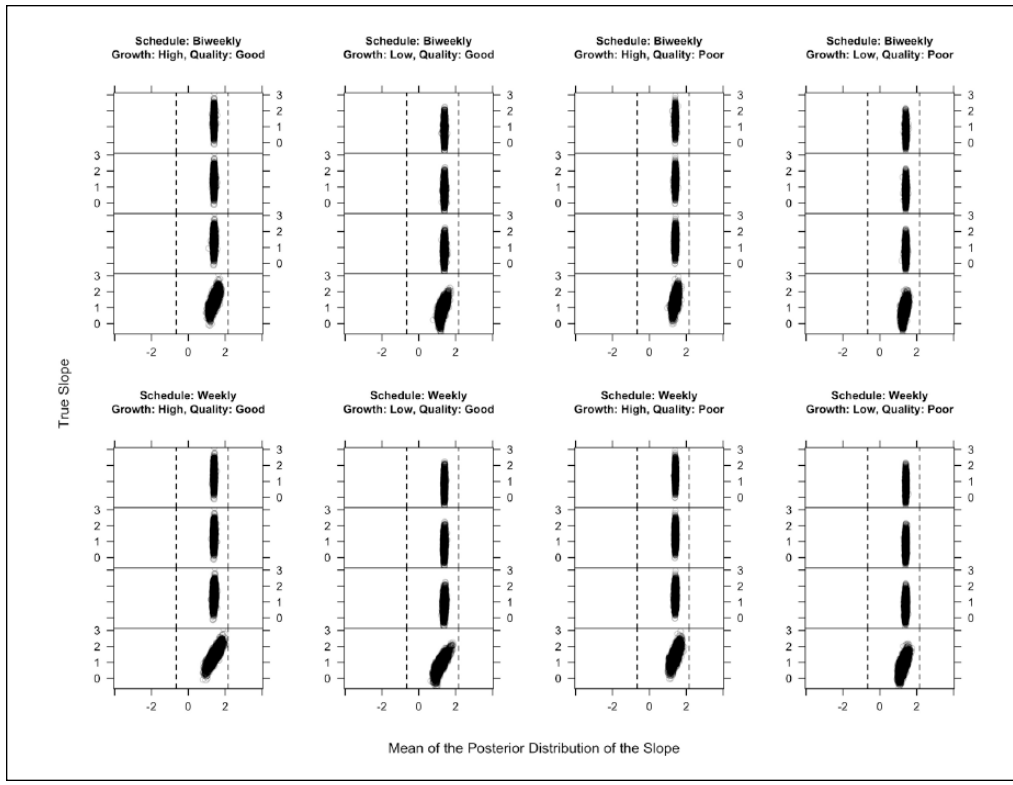

Figure 1. Scatterplot of true slope against the mean of the posterior distribution of the slope for the high growth slope (μ =1.4), low variance prior ($\sigma^2 = .2^2$).
Note. Each subplot is faceted by duration of progress monitoring (from top to bottom: 6, 8, 10, and 30 weeks of duration). The dotted vertical lines correspond to the minimum (left) and maximum (right) true score (right).

Descriptive Summary

Figures 1 through 5 are scatterplots that depict the relationships between true ROIs and posterior means of estimated ROIs (i.e., β_1 plotted against $\widehat{\beta}_1$) for priors of β_1 with low and high variance (the intermediate variance figures are available upon request). Figure 1 corresponds to the high prior mean for β_1 ($\mu = 1.4$) with the low variance prior ($\sigma^2 = .2^2$); Figure 2 corresponds to the low prior mean for β_1 ($\mu = 1.1$) with the low variance prior ($\sigma^2 = .2^2$); Figure 3 corresponds to the high prior mean for β_1 ($\mu = 1.4$) with the high variance prior ($\sigma^2 = .8^2$); and Figure 4 corresponds to the low prior mean for β_1 ($\mu = 1.1$) with the high variance prior ($\sigma^2 = .8^2$). Figure 5 is a scatterplot of the true ROI against estimated ROI for OLSR. Figure 6 presents a violin plot of the deviations of the true ROI from the estimated ROI (for OLSR) or the mean of the posterior distribution of the ROI (for the Bayesian models).

Reliability (and Validity)

A scatterplot of reliability against the duration of progress monitoring is presented in Figure 7. Recall that reliability was defined as the squared correlation of the true ROI and the estimated ROI (OLSR) or the posterior mean (Bayesian); therefore, visual analysis of Figures 1 to 5 also depicts results related to the reliability and validity of estimated ROIs.

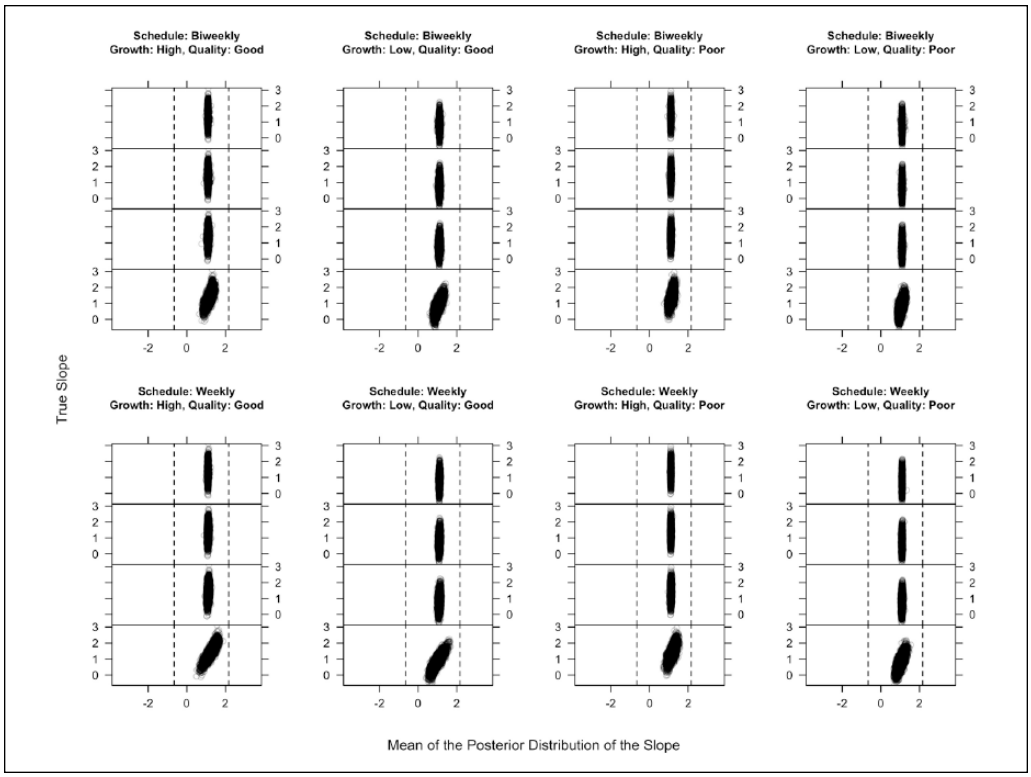

Figure 2. Scatterplot of true slope against the mean of the posterior distribution of the slope for the high growth slope (μ = 1.1), low variance prior ($\sigma^2 = .2^2$).
Note. Each subplot is faceted by duration of progress monitoring (from top to bottom: 6, 8, 10, and 30 weeks of duration). The dotted vertical lines correspond to the minimum (left) and maximum (right) true score (right).

MSD

Figure 8 shows a scatterplot of the MSD against the duration of progress monitoring by various conditions of the simulation study. Unsurprisingly, MSD was highly affected by the type of estimator and the duration of the progress monitoring. For all conditions, the MSD for OLSR fluctuated around zero, and as duration increased, OLSR approached zero.

RMSE

Figure 9 shows the RMSE against the duration of progress monitoring by various conditions of the simulation study. RMSE is a measure of the variability around the true ROI.

Regression

To examine the extent that the conditions of the simulation study differed in their variability around the true ROIs, we regressed the absolute value of the difference between the true ROI and the estimated ROI (for OLSR) or the posterior mean (for the Bayesian models) on the estimator type, level of student growth, and quality of the data for each combination of schedule and duration separately. We ran these analyses separately as it was clear from Figure 6 that homogeneity of variance would be violated because of the unequal number of observations (i.e.,

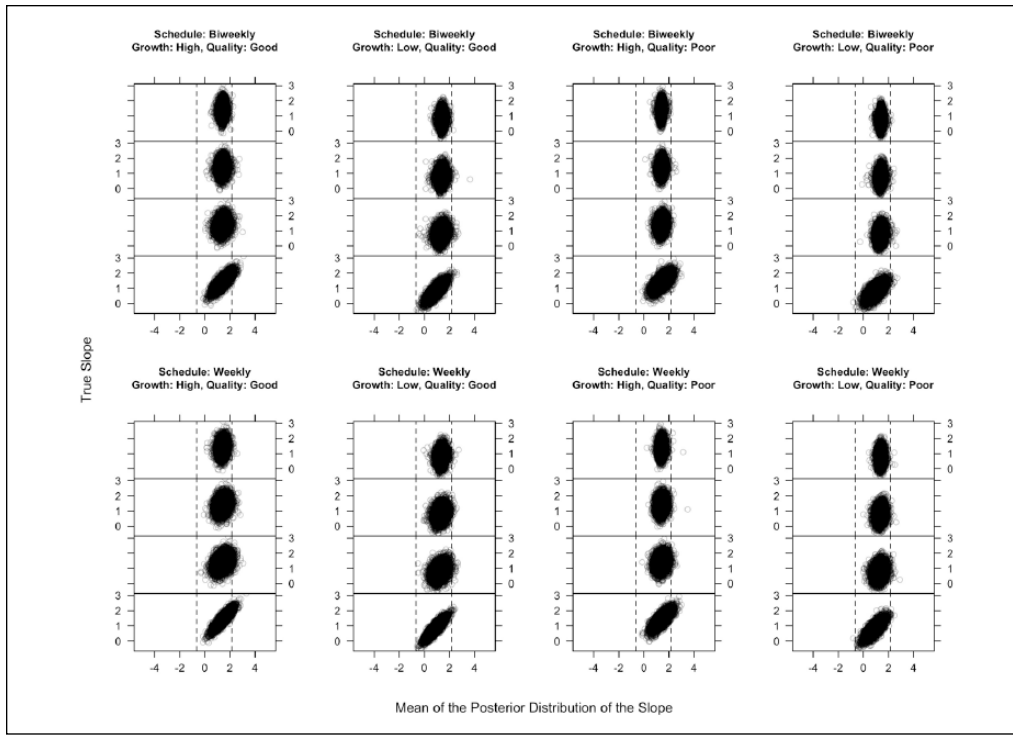

Figure 3. Scatterplot of true slope against the mean of the posterior distribution of the slope for the high growth slope (μ =1.4), low variance prior ($\sigma^2 = .8^2$).
Note. Each subplot is faceted by duration of progress monitoring (from top to bottom: 6, 8, 10, and 30 weeks of duration). The dotted vertical lines correspond to the minimum (left) and maximum (right) true score (right).

the variance would decrease as the number or measurements increased). However, even after running models based on the number of observations separately, we failed to have homogeneity of variance and used a log transformation to correct for this. This transformation corrected the nonconstant variance but resulted in negatively skewed residuals. However, given the agreement between Figure 6 and the results from the models, presented in Figure 9, the findings are likely robust to this observation.

All main effects of the estimator, level of student growth, or quality of the data, regardless of the level of duration or schedule considered, were statistically significant. Because of the large power from the large number of replicates of each condition, we instead examined and presented partial eta squared. Figure 10 presents a scatterplot of the partial eta squared for each model (a total of eight models were run, corresponding to a unique combination of duration and schedule). A column of points in Figure 10 corresponds to the partial eta squared for the estimator, student growth, and quality of data for a model based on a certain condition of duration and data collection schedule.

Discussion

The purpose of this research was to introduce the concept of Bayesian methods and compare them with OLSR to model CBM reading data for individual students. The use of Bayesian methods generally improved the estimates of ROIs (comparable MSDs and smaller RMSEs) and through shrinkage exerted by the prior distribution, resulted in ROIs that were plausible where

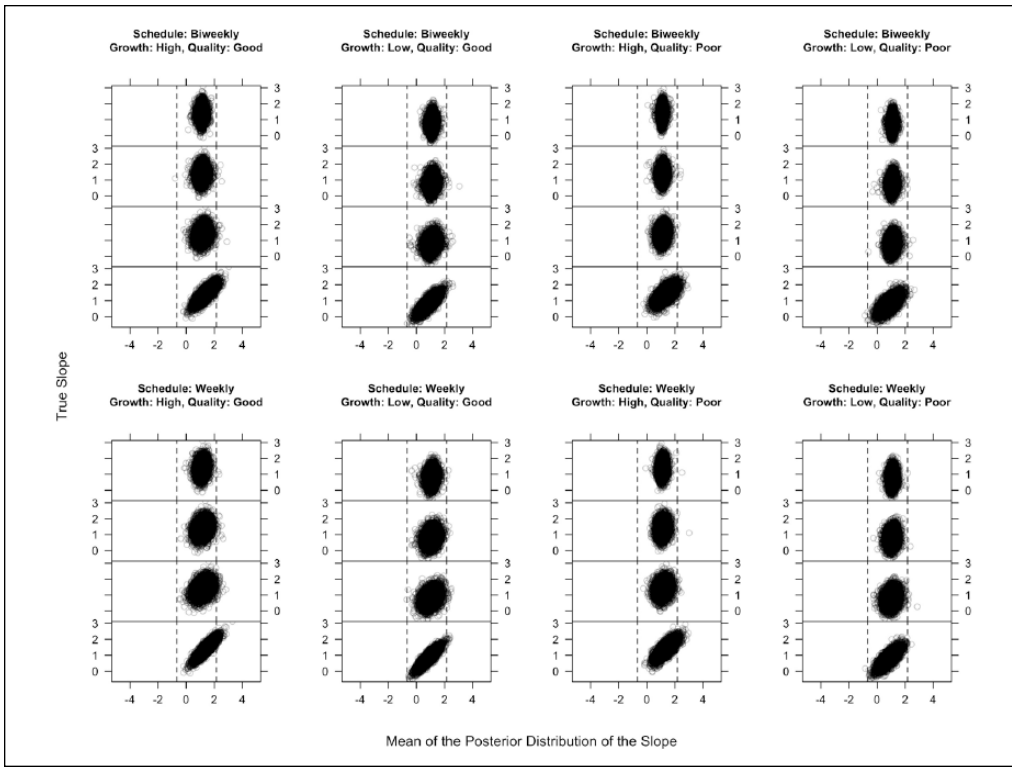

Figure 4. Scatterplot of true slope against the mean of the posterior distribution of the slope for the high growth slope (μ =1.1), low variance prior ($\sigma^2 = .8^2$).
Note. Each subplot is faceted by duration of progress monitoring (from top to bottom: 6, 8, 10, and 30 weeks of duration). The dotted vertical lines correspond to the minimum (left) and maximum (right) true score (right).

OLSR gave extreme estimates. These and other results from this study provide initial support for these applications of Bayesian methods.

Although the results are promising, they also illustrate that misspecified priors bias the estimated ROIs. The observed bias in estimated ROIs increased as the difference between the true parameter and Bayesian priors were increasingly more discrepant. That finding was expected. Nevertheless, those results are notable because they illustate that Bayesian methods do not represent a simple or generic solution. The specification of the Bayesian models and priors will substantially influence the results, which could confer benefit or harm if misused. Further exploration and experimentations are necessary to provide detailed specifications of the methods. The results of this study provide clear evidence that the priors used within the defined Bayesian model substantially influenced the reliability, validity, bias, and precision of ROI estimates. These results and implications briefly discussed below.

Precision of ROI Estimates

Precision improved as the duration of progress monitoring increased regardless of the method or model used to estimate ROIs (Figure 8); however, the Bayesian methods were more precise for all conditions with shorter durations of progress monitoring (6, 8, and 10 weeks). Precision converged for the two methods only for the extended duration (30 weeks). This is further illustrated in Figure 5, which depicts the distribution of deviations between

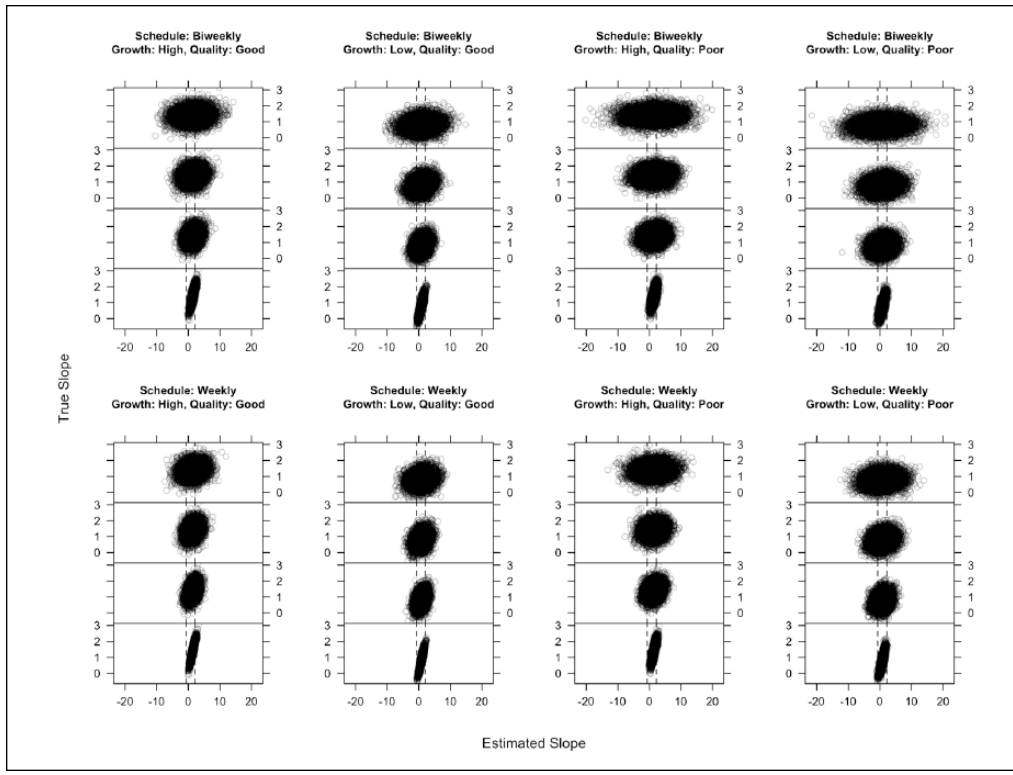

Figure 5. Scatterplot of true slope against the estimated mean from ordinary least squares regression.
Note. Each subplot corresponds to a level of schedule, student growth, and data quality, and each subplot is faceted by duration of progress monitoring (from top to bottom: 6, 8, 10, and 30 weeks of duration).

the estimated and true ROIs. This and other findings are further illustrated in Figures 1 to 4 and discussed below.

Relative effects. The choice to use either OLSR or Bayesian methods was by far the most influential factor on the precision of estimated ROIs (Figure 9). The effect was substantial during brief durations (6, 8, and 10 weeks) and more robust when data were collected less often (biweekly vs. weekly). These effects were substantially larger than those associated with the magnitude of student growth or the quality of data until Week 30, when all effects were similar.

Bayesian estimates. First, as the duration of the progress monitoring increases, the association between the true and the posterior mean of the ROI increases and becomes a strong, positive, linear relationship (Figures 1-4). That was consistent for both Bayesian and OLSR methods. Second, regardless of the prior distribution, the posterior means of the ROI posterior distributions were almost always within the limits of true growth (i.e., between the dashed horizontal lines); therefore, the prior distribution resulted in plausible ROI estimates (Figures 1-3). That was not the case for OLSR, which yielded a distribution with many implausible estimates (Figure 4).

Third, as the duration of the progress monitoring increases, the influence of the Bayesian prior diminishes (i.e., the shrinkage toward the prior mean decreases). This is most evident in Figure 1, which depicts the results for the more informative Bayesian prior (i.e., prior distribution with less variance). The estimated ROIs were distributed in a tight band in conditions with brief progress monitoring durations. The distributions resembled that of a vertical pipe with data clustered

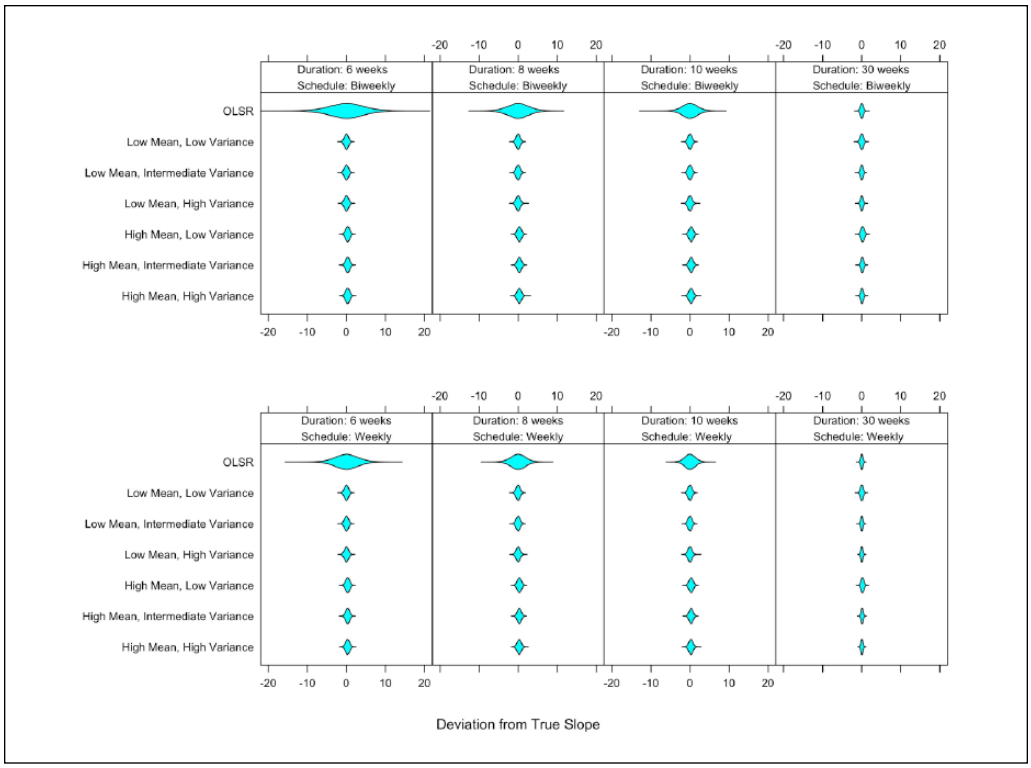

Figure 6. Violin plots of the deviation of the true slope from the estimate slope (for OLSR) or the mean of the posterior distribution of the slope (for the Bayesian models).
Note. OLSR = ordinary least squares regression.

tightly around the mean of the prior distribution. That band widened as the durations increased. The band also widened when the prior was less informative priors (compare Figures 1, 2, and 3).

Predictably, the priors influenced Bayesian estimates. Visual analysis of Figures 1 and 3 illustrates that result. The distributions of estimated ROIs were dramatically restricted with the more informative prior, especially for very brief durations of progress monitoring (Figure 1). In contrast, the distributions with the least informative prior covered near the full range of the true distribution—even for very brief durations (Figure 3).

OLSR estimates. The range of OLSR estimates included very unrealistic estimates of ROI when there were brief progress monitoring durations. That result was consistent across all conditions of growth, data quality, and schedule. Within some conditions, the approximate range of estimated ROIs spanned −20 to +20. That range far exceeds the range of true ROIs, which replicates previous findings (Christ, 2006; Christ, Zopluoglu, et al., 2013; Christ et al., 2012).

Reliability (and Validity)

The results depicted in Figures 1 to 4 also relate to the reliability and validity of the ROI estimates. An index of validity is the relationship/correlation between the true and observed score, and reliability is the square of this correlation. The coefficients for reliability also depicted in Figure 6, which demonstrates that reliability was generally similar for both Bayesian and OLSR methods across all conditions (Figure 6); however, less informative priors generally had the

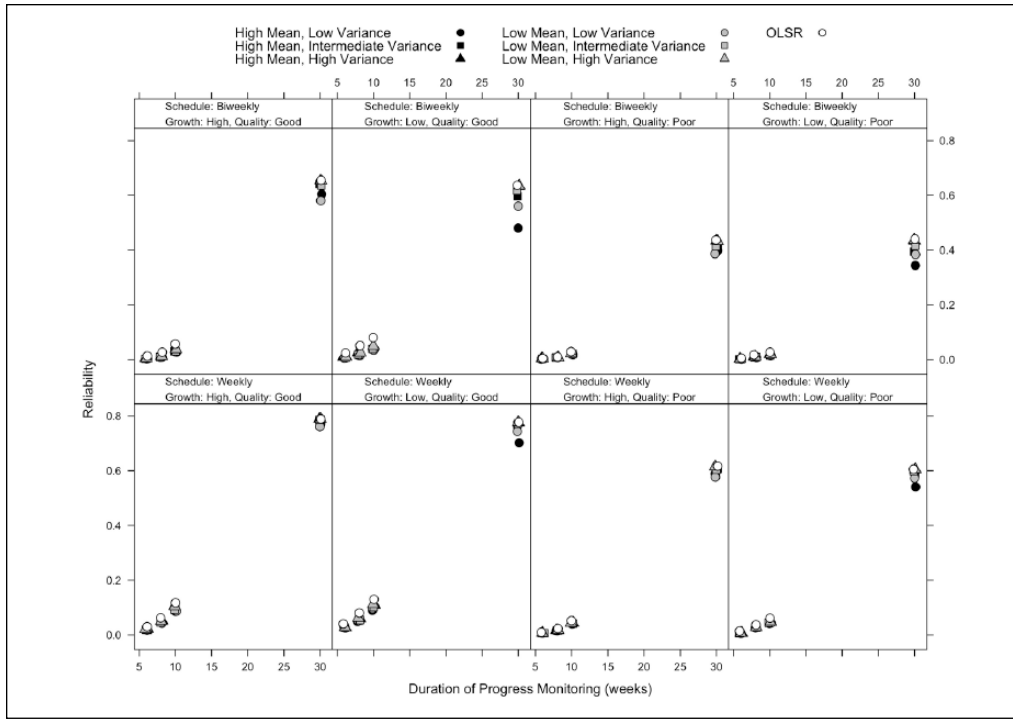

Figure 7. Scatterplot of reliability against duration of progress monitoring by estimator, schedule, student growth, and quality of data.

Note. The top row corresponds to collecting data weekly and the second row corresponds to collecting data biweekly. The first two columns correspond to high-quality CBM-R data and the last two columns correspond to poor-quality CBM-R data. Finally, the first and third columns correspond to the true high CBM-R growth, and the second and fourth columns correspond to true low CBM-R growth. A small horizontal jitter was used. OLSR = ordinary least squares regression; CBM-R = Curriculum-Based Measurement of Oral Reading.

lower reliabilities and OLSR usually had the highest reliabilities. Higher quality data, weekly schedules, and longer durations improved reliability; however, it was only at 30 weeks of duration that strong, positive, and linear relationships between true and estimated ROIs emerged (Figures 1-4).

Bias of ROI Estimates

OLSR provided unbiased estimates of ROI. The Bayesian method often provided estimates with either positive or negative bias, which depended on the convergence between the priors used to estimate ROIs and true distribution of ROIs. Bias was most evident and robust when there were larger discrepancies between the true mean and the specified prior; especially when the prior was more informative (i.e., less variance). Regardless, that bias generally decreased as the duration increased.

The pattern of bias is illustrated in Figure 7. The first and third columns in the first plot depict bias for conditions with high true ROI (1.4 WRCM per week). In that condition, high mean prior distribution (1.4 WRCM per week) matched the mean of the true ROI, which is why there was essentially no bias regardless of duration, schedule, or quality. In contrast, the low mean prior distribution (1.1 WRCM per week) resulted in biased estimates (approximate range = −0.2 to −0.3 WRCM per week).

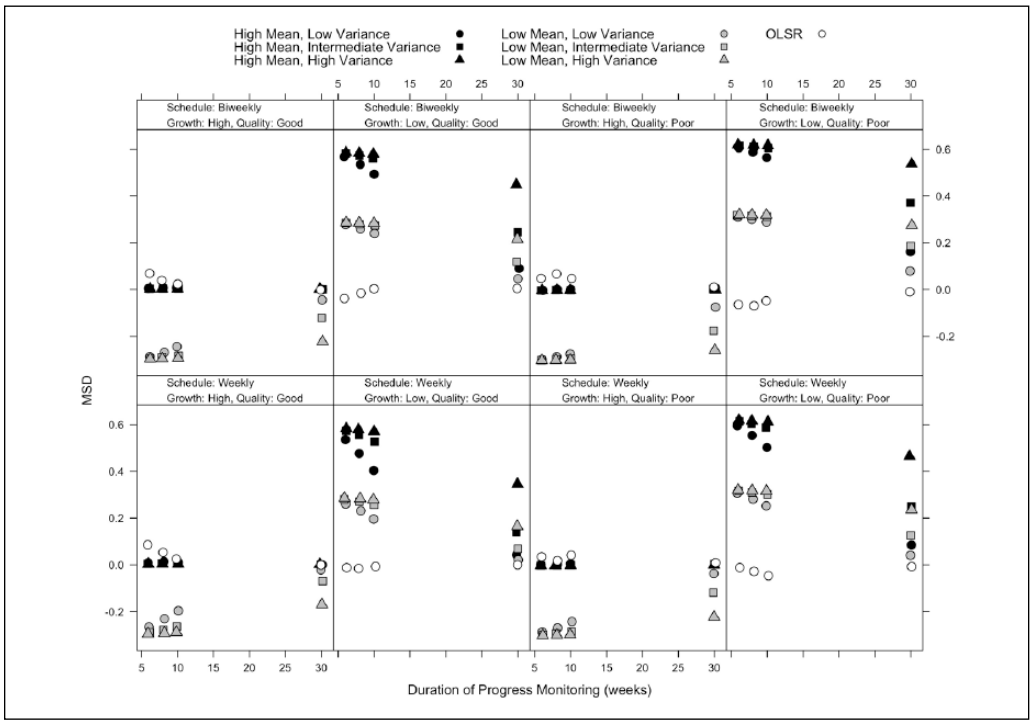

Figure 8. Scatterplot of MSD against duration of progress monitoring by estimator, schedule, student growth, and quality of data.

Note. The top row corresponds to collecting data weekly and the second row corresponds to collecting data biweekly. The first two columns correspond to high-quality CBM-R data and the last two columns correspond to poor-quality CBM-R data. Finally, the first and third columns correspond to the true high CBM-R growth, and the second and fourth columns correspond to true low CBM-R growth. A small horizontal jitter was used. OLSR = ordinary least squares regression; MSD = mean signed difference; CBM-R = Curriculum-Based Measurement of Oral Reading.

Conclusion

The results of this study illustrate that Bayesian methods can be used to estimate ROIs; however, there are important considerations. The most important consideration might relate to the specification of the prior and the model. Only one Bayesian model was examined here, but a range of priors were specified to illustrate their influence. Although other Bayesian models should be explored, the results this study clearly illustrate that the specification of priors is an important consideration. The values selected here were generally consistent with previously published research (Christ, 2006; Christ, Zopluoglu, et al., 2013; Christ et al., 2012; Deno et al., 2001). Although other priors might be examined in future research, it is very clear that misspecified prior means will bias estimates and more informative priors (i.e., less variance) will amplify that effect. Well-specified means and modestly informative priors provided the most improvement in the estimation of ROIs, which were more precise and unbiased.

Although OLSR was consistently unbiased and had similar reliability to Bayesian estimates, the precision of those estimates was very poor after brief period of progress monitoring (6-8 weeks). The range of true values approximated 0 to 3 WRCM per week. The range of estimated values with OLSR was −20 to 20 WRCM per week in some conditions. OLSR is simply insufficient. This study provides preliminary methods and evidence to support the ongoing research and development to improve the estimation of ROIs using alternate methods.

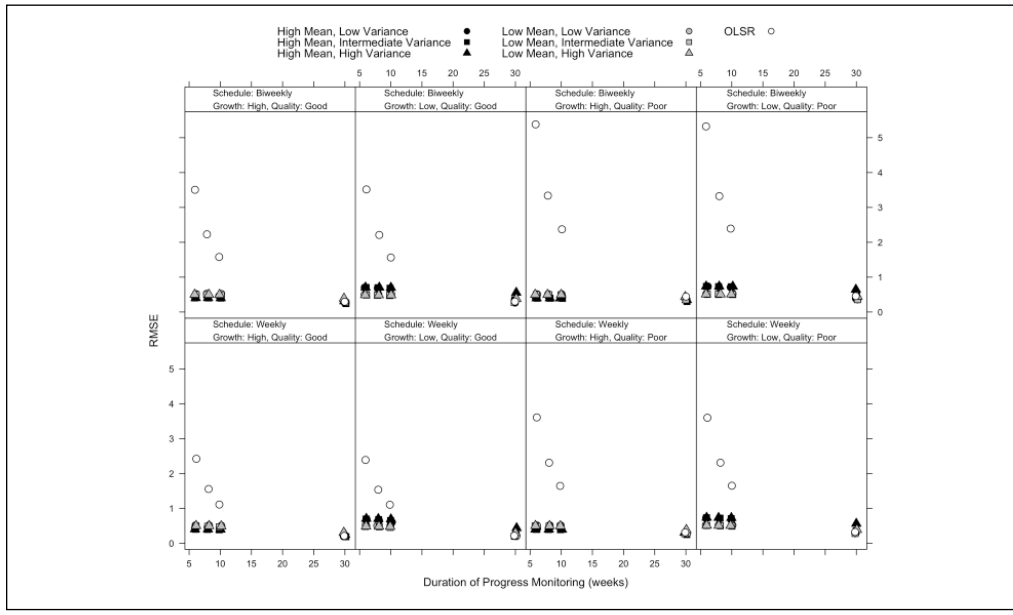

Figure 9. Scatterplot of RMSE against duration of progress monitoring by estimator, schedule, student growth, and quality of data.
Note. The top row corresponds to collecting data weekly and the second row corresponds to collecting data biweekly. The first two columns correspond to high-quality CBM-R data and the last two columns correspond to poor-quality CBM-R data. Finally, the first and third columns correspond to the true high CBM-R growth, and the second and fourth columns correspond to true low CBM-R growth. A small horizontal jitter was used. RMSE = root mean square error; OLSR = ordinary least squares regression; CBM-R = Curriculum-Based Measurement of Oral Reading.

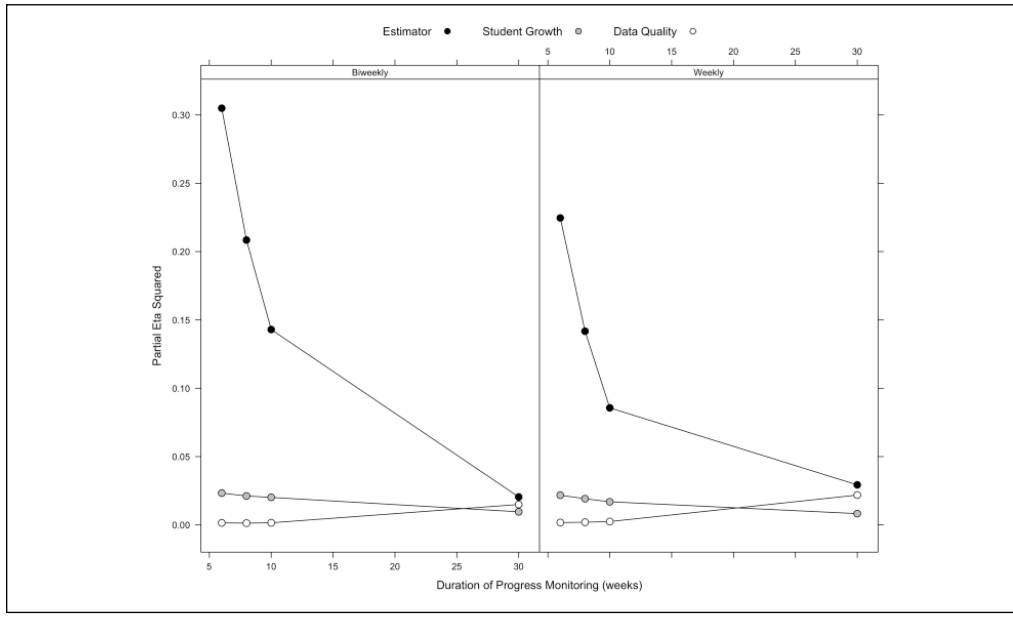

Figure 10. Scatterplot of partial eta squared against duration of progress monitoring by estimator, schedule, student growth, and quality of data, and faceted by data collection schedule.
Note. Each column of points corresponds to a model with a unique number of assessments administered (i.e., a combination of schedule and duration of progress monitoring). From left to right (in both subplots), the points correspond to partial eta squared from a model with 6, 8, 10, and 30 weeks of duration. The panels depicts biweekly (left) and weekly (right) data collection schedules.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Theodore J. Christ, PhD, has equity and royalty interests in, and will serve on the Board of Directors for, FastBridge Learning (FBL) LLC, a company involved in the commercialization of the Formative Assessment System for Teachers (FAST). The University of Minnesota also has equity and royalty interests in FBL LLC. These interests have been reviewed and managed by the University of Minnesota in accordance with its Conflict of Interest policies.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324A130161 to the University of Minnesota.

References

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

Ardoin, S. P., Christ, T. J., Morena, L. S., Cormier, D. C., & Klingbeil, D. A. (2013). A systematic review and summarization of the recommendations and research surrounding Curriculum-Based Measurement of Oral Reading Fluency (CBM-R) decision rules. *Journal of School Psychology*, *51*, 1-18. doi:10.1016/j.jsp.2012.09.04

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

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., Monaghen, B., Zopluoglu, C., & Van Norman, E. R. (2013). Curriculum-Based Measurement of Oral Reading (CBM-R): Evaluation of growth estimates derived with pre-post assessment methods. *Assessment for Effective Intervention*, *37*(4), 19-57. doi:10.1177/1534508412456417

Christ, T. J., & Vining, O. (2006). Curriculum based measurement procedures to develop multiple-skill mathematics computation probes: Evaluation of random and stratified stimulus-set arrangements. *School Psychology Review*, *35*, 387-400.

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., & Van Norman, E. R. (2013). Curriculum-Based Measurement of Oral Reading (CBM-R) progress monitoring: Multi-study evaluation of schedule, duration, and dataset quality on progress monitoring outcomes. *Journal of School Psychology*, *51*, 19-57. doi:10.1016/j.jsp.2012.11.001

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 Journal of Special Education*, *24*, 160-173.

Deno, S. L. (1995). The school psychologist as problem solver. In J. Grimes & A. Thomas (Eds.), *Best practices in school psychology III* (pp. 37-56). Bethesda, MD: 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 measurement to establish growth standards for students with learning disabilities. *School Psychology Review*, *30*, 507-524.

Deno, S. L., Marston, D., & Tindal, G. (1985). Direct and frequent curriculum-based measurement: An alternative for educational decision making. *Special Services in the Schools*, *2*(2), 5-27.

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

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology*, *46*, 315-342.

Fuchs, D., Fuchs, L. S., McMaster, K. N., & Al Otaiba, S. (2003). Identifying children at risk for reading failure: Curriculum-based measurement and the dual-discrepancy approach. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 431-449). New York, NY: Guilford Press.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, *57*, 488-500.

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Good, R. H., & Shinn, M. R. (1990). Forecasting accuracy of slope estimates for reading curriculum-based measurement: Empirical evidence. *Behavioral Assessment*, *12*(2), 179-193.

Gresham, F. (2001). *Responsiveness to intervention: An alternative to the identification of learning disabilities*. Paper presented at the 2001 Learning Disabilities Summit: Building a Foundation for the Future. Retrieved from http://ldsummit.air.org/paper.htm

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

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

Individuals With Disabilities Education Improvement Act, 20 U.S.C., Pub. L. No. 108-446 § 1400 *et seq*. (2004).

Jenkins, J. R., Graff, J. J., & Miglioretti, D. L. (2009). Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children*, *75*(2), 151-163.

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

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam: Academic Press.

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.

Shinn, M. R., Good, R. H., & Stein, S. (1989). Summarizing trend in student achievement: A comparison of methods. *School Psychology Review*, *18*(3), 356-370.

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice*, *15*, 128-134.

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.

Thornblad, S. C., & Christ, T. J. (2014). Curriculum-based measurement of reading: Is 6 weeks of daily progress monitoring enough? *School Psychology Review*, *43*(1), 19-29.

Van Norman, E. R., & Christ, T. J. (2016). Curriculum-based measurement of reading: Accuracy of recommendations from three-point decision rules. *School Psychology Review*, *45*, 296-309.

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.

University of Oregon Center on Teaching and Learning. (2017). *Office DIEBELS home page: UO DIEBELS data system*. Retrieved from https://dibels.uoregon.edu.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, *41*(2), 390-420.