The Effects of Baseline Estimation on the Reliability, Validity, and Precision of CBM-R Growth Estimates

Ethan R. Van Norman, Theodore J. Christ, and Cengiz Zopluoglu
University of Minnesota, Twin Cities

This study examined the effect of baseline estimation on the quality of trend estimates derived from Curriculum Based Measurement of Oral Reading (CBM-R) progress monitoring data. The authors used a linear mixed effects regression (LMER) model to simulate progress monitoring data for schedules ranging from 6–20 weeks for datasets with high and low levels of residual variance (poor and good quality datasets respectively). Three observations per day for the first three days of data collection were generated for baseline estimation. As few as one and as many as nine observations were used to calculate baseline. The number of weeks of progress monitoring and the quality of the dataset were highly influential on the reliability, validity, and precision of simulated growth estimates. Results supported the use of using the median of three observations collected on the first day to estimate baseline, particularly when the first observation of that day systematically underestimated student performance. Collecting a large number of observations to estimate baseline does not appear to improve the quality of CBM-R growth estimates.

Keywords: curriculum based measurement, oral reading fluency, simulation, baseline

Educators and researchers developed Curriculum Based Measurement (CBM) to assess student performance in relation to instruction in basic academic skill areas frequently, reliably, and efficiently (Deno, 1985). CBM uses standardized testing procedures, is sensitive to short-term improvement, and requires minimal training to administer and score (Shinn, 2008). CBM enables educators to determine the degree to which academic interventions improve student performance on an ongoing basis and indicates when instructional modifications are necessary. There are many forms of CBM (e.g., spelling, written expression, and math computation), but oral reading fluency (CBM-R) is the most researched and widely used (Wayman, Wallace, Wiley, Ticha, & Espin, 2007).

Reading competence is fundamental to academic success and oral reading fluency is a robust indicator of broad reading competence (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Reschly, Busch, Betts, Deno, & Long, 2009; Shinn, Good, Knutson, & Collins, 1992; Wayman et al., 2007; Yeo, 2010). Calculating a slope through repeated measures of words read correct per minute (WRCM) out loud, across time, quantifies a rate of change in oral reading fluency. The professional literature suggests that a minimum of 10 observations should be collected before calculating a slope (Good & Shinn, 1990; Shinn, 2002; Shinn, Good, & Stein, 1989). Christ (2006) suggested that two observations be collected per week for 10 weeks.

Calculating the rate of change of WRCM in the presence of instruction should provide an index of how effective that instruction is at improving reading competence (Deno, 1985;
Deno, 1986; Deno, Marston, & Tindal, 1985). If CBM-R were a perfect index of growth relative to instruction, instruction would be the only source of variability in WRCM across observations (Jenkins, Zumeta, Dupree, & Johnson, 2005). That is, a rate of improvement in WRCM, or a slope value, would indicate how effective instruction, and nothing else, is at improving reading competence. In addition, estimates of slope would be the same regardless of calculation methods (e.g., moving median vs. ordinary least squares), the duration (number of weeks) or density (number of observations per week) of data collection.

Sources of Instability

Differences in WRCM across observations are not solely indicative of improvement in reading competence. In fact, WRCM from alternate forms, administered within a short amount of time, can differ by as many as 40 words (Christ & Ardoin, 2009). In addition, the setting of administrations (Derr-Minneci, 1990), delivery of instructions (Colón & Kanzler, 2006), and level of text difficulty (Betts, Pickart, & Heistad, 2009; Christ & Ardoin, 2009; Hintze & Christ, 2004; Hintze, Daly, & Shapiro, 1998; Poncy, Skinner, & Axtell, 2005) are known sources of instability in WRCM estimates. The quality of a CBM-R dataset is the product of psychometric properties of the instrument (e.g., passage equivalence) the consistency of administration procedures, and perhaps the day-to-day idiosyncratic characteristics of the data collector and student. In other words, there may be sources of error associated with CBM-R that cannot be explained by instrumention or administration procedures alone. For instance, a student may have a lower level of motivation or just may be having an “off-day” on any particular administration. Researchers need to investigate the extent, if any, such day-to-day events introduce error into CBM-R estimates.

Dataset quality subsumes all validated and hypothesized sources of error and is operationalized as the residual variance, or the magnitude of error or “bounce” of WRCM observations around a computed trend line. The level of bounce has been conceptualized as the standard error of the estimate in previous literature (Fuchs, Fuchs, & Deno, 1982; Poncy et al., 2005; Shinn, Gleason & Tindal, 1989). High-quality dataset estimates are tightly grouped, and low-quality dataset estimates are highly spread around a trend line. Christ, Zopluoglu, Long, and Monaghan (2012), defined very good, good, poor, and very poor quality datasets as having residual variance (σε), or magnitude of error, equal to 5, 10, 15, and 20, respectively. Those definitions converge with analysis of several empirical studies (e.g., Ardoin & Christ, 2009; Francis et al., 2008; Good & Shinn, 1990; Jenkins et al., 2005; Shinn, Gleason, & Tindal, 1989). Aside from optimizing the instrumentation (i.e., psychometric properties) and administration procedures (the process of collecting and scoring probes) of CBM-R, optimizing methods to calculate slope may minimize the average level of spread around a computed trend line, and limit the impact of dataset quality when estimating growth.

Research on optimizing trend estimation has generally focused on three questions: (1) How should we calculate slope (Fuchs & Fuchs, 1986; Deno, Fuchs, Marston, & Shinn, 2001; Parker & Tindal, 1992; Shinn, Good, & Stein, 1989), (2) how long or how often should we collect data (Jenkins, Graff, & Miglioretti, 2009; Jenkins, Hudson, & Lee, 2007; Jenkins & Terjesen, 2011), and (3) how many observations should we collect at a time (Fuchs, 2003; Jenkins et al., 2009)? Ordinary least squares (OLS) regression is the preferred method of slope calculation, but there are disagreements regarding the optimal duration and density of data collection (Ardoin, Christ, Morena, Cormier, Klingbeil, 2013). This study introduces a fourth question pertinent to optimizing growth estimates: How should we estimate the initial level of performance, or baseline, upon the initiation of progress monitoring?

Baseline Estimation

Other areas of educational research, particularly applied behavior analysis and behavioral consultation, emphasize the importance of baseline estimation. Research in both fields highlight the need for stable and accurate initial observations of behavior. Clinicians determine the necessity of a treatment plan based on initial frequencies of behavior (Kratochwill & Bergan, 1990). In addition, intervention effectiveness is evaluated by comparing the rate of behavior
during an intervention phase against a baseline or preintervention phase (Kazdin, 2011; Ximenes, Manolov, Solanas, & Quera, 2009). Inaccurate or unstable baseline estimates, typically when there is a presence of trend during the baseline phase, introduces uncertainty in the determination for the need of treatment and the evaluation of intervention effects (Riley-Tillman, Burns, & Gibbons, 2013). Researchers that use single case design methodologies have attempted to find ways to optimize baseline estimation to avoid such problems (Furlong & Wampold, 1982; Parker, Cryer, & Byrnes, 2006; Scruggs & Mastropieri, 1998).

Several commercially available software programs can estimate a slope value and draw a trend line through CBM-R data with a few key strokes. Technological advancements combined with the sheer number of students being monitored may cause educators to rely less and less on informal methods to discern trend when evaluating CBM-R data. As a result, there may be less of an emphasis on collecting an adequate number of observations to estimate baseline (Hixson, Christ, & Bradley-Johnson, 2008). The presence of trend when calculating baseline for CBM-R is not a great concern because within week growth is typically minimal (Deno et al., 2001). Instructional documents for estimating baseline for CBM-R typically focus on deriving a precise estimate of level and suggest the administration of three probes on the first day of data collection using the median value as a baseline estimate (Shinn & Shinn, 2002). Using a median value as a baseline estimate has many advantages. Median values are less susceptible to influence from extreme values and are easy to compute. But it is unclear whether using the median observation of three observations is the optimal method to estimate baseline performance in the context of CBM-R. That is, can improving the precision of level in baseline estimation improve estimates of growth for CBM-R?

Jenkins and colleagues (2009) evaluated the validity of growth estimates after manipulating the number of observations to estimate baseline. True slopes were calculated for 41 students from 29 probes administered across 10 weeks. Validity was defined by the correspondence of true slope to estimated slope. The former was defined by the full dataset, and the latter was calculated from observations from either every two, three, four, or nine weeks. Across schedules, slope estimates using the median of multiple observations yielded more valid results than slopes using the first observation as baseline. That is, using the first observation as baseline systematically overestimated slope. This finding indicated that the first observation of the first day systematically underestimated performance. The researchers recommended to establish baseline, four passages be administered. Educators should discard the first observation, and use the median of the three remaining observations.

The finding of Jenkins et al. (2009) provides initial evidence for using multiple observations to estimate baseline, and coincide with some recommendations in the best practices literature (Riley-Tillman & Burns, 2010; Shinn, 2008). Jenkins and colleagues concluded that it was inadvisable to use the first observation as a baseline estimate because it biased slope estimates. The researchers did not investigate whether bias was present when data were collected for a longer amount of time. In addition, they did not determine if bias existed when estimating slope with different quality datasets. This study builds upon the findings of Jenkins et al. by using simulation methods to evaluate how different baseline estimation procedures affect estimates of slope while also considering the duration of data collection and quality of datasets.

Simulation Methodology

Simulation methods are frequently used in other areas of educational assessment. For instance, within computer adaptive testing, simulation methods are used to determine the ideal number of items that should be administered before a test ends (Weiss & Guyer, 2010). As a point of comparison, it is useful to know the level of error, or precision, of a student’s ability estimate if the student took every item in an item bank. Researchers then weigh the cost of administration time against the benefits of precision of different length tests. If an item bank consists of thousands of items, and the content of those items are highly sensitive (e.g., state achievement tests), is it truly feasible to administer all questions to see how precise of an ability estimate is possible? Further, is it possible to administer a 20, 30, and 40 item test and
control for threats to internal validity (e.g., practice effects, maturation, and fatigue) when researching with actual students? Lastly, many of the datasets used for calibrating items within a computer adaptive test rely on sampling plans where students do not answer every item in the item bank (i.e., common-item nonequivalent groups design; Kolen & Brennan, 2010). Instead, researchers use simulation methods to model student responses to items they did not answer based on the response pattern to items they did answer (Nydick & Weiss, 2009). After that, researchers simulate different length adaptive tests and levels of precision across different ability estimates are compared. After initial simulations, researchers confirm their findings via post hoc analysis of field-based data (e.g., Van Norman & Christ, 2012). Virtually every computer adaptive test used in schools is developed and refined using simulation methodology (De Ayala, 2009).

Simulation methods offer a starting point to investigate the relationship between several facets of data collection procedures and progress monitoring outcomes simultaneously that are infeasible if not impossible to study via a single field-based data collection. Instead of spending large amounts of money, and causing students to miss instructional time for research purposes, Christ et al. (2013) contend that it is reasonable, as a first step, to investigate different progress monitoring procedures by (a) beginning with a large extant database, (b) cross referencing model parameters from various published studies, and (c) conducting simulations to supplement the available findings. The results of simulations come with a high level of internal validity and offer an effective, efficient, and economical starting point to investigate different progress monitoring practices in applied practice. Justification for using simulation methodology as an initial low risk way of investigating applied problems in fields as diverse as agriculture, meteorology, medicine, and economics follows a similar cost-benefit argument (Robinson, 2004).

Previous CBM-R Simulation Research

Christ et al. (2012) used simulation methods to evaluate how slope calculation methods, quality of datasets, and number of observations affected estimates of growth. Similar to this study, the researchers first analyzed a large high quality extant CBM-R database ($n = 3,078$ AIMSweb progress monitoring cases). From initial analyses, parameters for a data generation model were calculated, data for hypothetical individuals were generated, and true and observed slopes computed. In their study, the researchers evaluated a data collection schedule in which one observation was collected per week. The researchers found that slope calculation method, dataset quality, and the number of observations collected affected the reliability and precision of slope estimates. The researchers concluded that to use CBM-R growth estimates to make low stakes decisions—such as making alterations to instructional design (Thorndike & Thorndike-Christ, 2010) - indicated by reliability coefficients of at least .70 (Kelly, 1927; Salvia, Ysseldyke, & Bolt, 2009), educators should (a) calculate slope using OLS regression, (b) use instruments and administration procedures that yield very good (high quality) datasets ($e = 5$), and (c) collect a minimum of 14 observations. For good quality data sets ($e = 10$), almost 20 observations are required to make low stakes decisions.

Christ and colleagues (2013) extended the findings of Christ et al. (2012) to evaluate how the density and durations of data collection schedules affected estimates of growth. The researchers sought to determine whether collecting more observations each week could yield more reliable and precise estimates of growth in a shorter amount of time. In other words, the researchers investigated whether low stakes decisions were feasible sooner if more data were collected each week. The primary finding of that study was that the duration of progress monitoring schedules accounted for the most variation in the precision of slope estimates. The authors concluded that instructional effects need time to substantiate. The researchers also found that the density of data collection accounted for a significant amount of variation in the precision of slope estimates. By collecting three observations per week with a good quality data set, low stakes decisions could be made after 15 weeks, nearly 5 weeks sooner than collecting one observation per week. One finding that was not clear from the study was whether dense data collection was required throughout the duration of progress monitoring schedules.
Purpose

The purpose of this study is to determine the effects of baseline values on estimates of trend in CBM-R data. This study builds upon the findings of Jenkins et al. (2009) by using simulation methods similar to Christ et al. (2012) and Christ et al. (2013). The work was guided by two research questions: How does baseline estimation affect the reliability, validity, and precision of estimates of growth, and what, if any, value is added by collecting a large number of observations to calculate baseline? We hypothesized baseline estimation procedures that used the mean or median of three days of observations would improve the reliability, validity, and precision of growth estimates compared with using a single or the median of three observations.

The answers to these questions have direct implications for educational decision making. Improving the capability for CBM to capture student responsiveness will ultimately increase the feasibility of its use in determining intervention effectiveness and informing decisions in a problem solving model (Shinn, 2008). In fact, outcomes of progress monitoring are being used to not only make instructional modifications, but serve as important information for placement decisions (Marston, Muyskens, Lau, & Carter, 2003; Shinn, 2007). Improving the precision of a baseline estimate may in turn increase the precision of growth estimates and ultimately improve the sensitivity to which CBM captures student responsiveness. Improving the quality of progress monitoring estimates could increase the defensibility of using CBM progress monitoring data to make high stakes decisions. As a result, it behooves educators and psychologists to evaluate every facet of progress monitoring practices to increase the reliability, validity, and precision of growth estimates.

In addition, the potential improvement in progress monitoring outcomes needs to be weighed against the time money and resources needed to collect that data (Hixson et al., 2008). This balance is especially relevant to consider because one of the major criticisms of multitiered systems of support, which CBM-R progress monitoring is frequently used in, is the amount of resources being allocated to assessment activities (Wixson & Valencia, 2011).

The results of this study could not only underscore the need for a stable baseline estimate when initiating progress monitoring, but may also provide evidence for the current practice of using the median of three observations as baseline both from a psychometric and cost–benefit standpoint. Alternatively, the results may serve as impetus to evaluate other baseline estimation strategies and force educators and school psychologists to perform a cost–benefit analysis of using those baseline estimation methods.

Method

Statistical Model for Data Generation

Longitudinal modeling is used to evaluate data from repeated measures across time. In this study, longitudinal modeling was used to examine the rate of change in oral reading fluency by evaluating the relationship between time and WRCT. Linear mixed effects regression (LMER) was used to model hypothetical repeated measures for individuals via the equation used by Christ et al. (2013):

\[
Y_{ijk} = (\beta_0 + b_{0k}) + (\beta_1 + b_{1i})Time_{ij} + \varepsilon_{ijk}. \tag{1}
\]

Where \(Y_{ijk}\) was the \((k = 1, 2, 3)\) observed level of performance for individual \(i (i = 1 \ldots n)\) at time point \(j (j = 1 \ldots n)\). The terms \(\beta_0\) and \(\beta_1\) are the group level regression coefficients for intercept and slope, respectively. These two terms are fixed effects. The terms \(b_{0i}\) and \(b_{1i}\) are the hypothetical individual’s deviations of intercept and slope (indicated by the \(i\) subscript) from the group level estimates. These terms are known as random effects. The mixture of fixed and random effects is where the term mixed effects regression is derived from. The last term \(\varepsilon_{ijk}\) represents the residual or random error of the \(k\)th observation for the \(i\)th individual at the \(j\)th time point.

True slope was estimated from the fixed and random effect terms \((\beta_1 + b_{1i})\)’Time_{ij}\) for each individual. A random value from a prespecified distribution of residual terms \(\varepsilon_{ijk}\) was added to each observation, and an OLS regression line was fitted to estimate observed slope for each individual.
Model Assumptions for Data Generation

Common assumptions associated with LMER (Fitzmaurice, Laird, & Ware, 2011; Long, 2012, p. 168; Verbeke & Molenberghs, 2000) were observed when generating data. First, random effects had a joint-normal distribution with means of 0 and variance-covariance matrix $G$. The variance of the random intercepts and the random slopes are denoted as $Var(b_{0j})$ and $Var(b_{1j})$ respectively, with the covariance between them $Cov(b_{0j}, b_{1j})$. Second, random errors had a multivariate normal distribution with means of 0 and variance-covariance matrix $R$ with constant variances ($\sigma^2$) on the diagonal. Finally, the random errors and random effects were uncorrelated.

The $k$ subscript of equation 1 modeled three CBM-R data points per observation for the first three days of week 1. The random errors associated with the $k$ observations at time point $j$ for person $i$ was assumed to be correlated. Across observations errors were assumed to be uncorrelated. Both assumptions are founded in psychometric research and theory (Long, 2012, p. 168).

Deriving Parameter Values for Data Generation

The data collection schedule represented a scenario in which three CBM-R probes were administered each day for the first three days of a school week, followed by the administration of one probe per week for the duration of progress monitoring. Three days were chosen (as opposed to four or five) to keep analyses manageable and better reflect what would likely occur in schools. To simulate data, values for the parameters of equation 1 had to be derived. Based on an analysis of a large extant database, described below, and expert opinion (from researchers who have published extensively in the fields of educational measurement and CBM), responses were estimated with known values for $b_0, b_1, Var(b_{0i}), Var(b_{1i}), Cov(b_{0i}, b_{1i}), \sigma^2, \rho_{01}, p$. To elaborate, expert opinion was gathered from four doctorate-level researchers in educational psychology across four states. All researchers had performed large-scale federally funded research investigating the psychometric properties of CBM-R. In addition, two professors studying educational measurement across two states were consulted. Both have published multiple peer-reviewed articles as well as textbooks on simulation methodology and longitudinal modeling. Of the parameters estimated, the level of residual variance $\sigma^2$ differed among simulations. All other parameters were constant values. The simulations were carried out with the R computer program (R Development Core Team, 2009).

Parameter estimates were derived from the same extant dataset as previous CBM-R simulation studies (Christ et al., 2012, 2013). LMER models were fitted to a dataset of second ($n = 1517$) and third ($n = 1561$) grade students. The demographic distribution of the sample was nearly 46% Female, 2% Special Education, 53% White, 17% Black, 8% Hispanic/Latino, 6% Asian/Pacific Islander, and 2% American Indian/Alaska Native across grades. The data were collected as part of a federally funded project designed to provide supplemental (Tier II) reading interventions to students at risk of reading difficulties. Interscorer reliability data were not available, but data collectors were trained to criterion with AIMSweb training materials and assessed for administration fidelity with the Accuracy of Implementation Rating Scales (Shinn & Shinn, 2002).

Table 1 summarizes the parameter estimates. The average number of WRCM across both grade levels was 40 with a standard deviation of 12.2 ($var = 150$). From these results, the fixed effect for group intercept ($b_0$) was set at 40 and the variance for random intercepts, $Var(b_{0i})$, was 150. The weekly slope estimate approximated 1.50 WRCM with standard deviation of .63 ($var = .40$). Baseline data collection required estimates for daily growth, so the average daily slope estimate was assumed to be $1.5/7 = .21$ WRCM. The corresponding standard deviation being $.63/7 = .09$ ($var = .0081$). The fixed effect for group slope was set at $b_1 = .21$, and the variance of the random slopes, $var(b_{1i}) = .0081$ in the model across conditions for simulations. The correlation between random effects was approximately .20.

For the baseline phase of data generation, the correlation between errors associated with $k$ successive observations was fixed at $.10$. This value was observed in another large extant data set where 500 students were administrated 65 passages within 10 days with as many as 10
passages administered on an occasion (Christ, Ardoin, Eckert, & White, 2010).

Independent Variables

Four independent variables were manipulated in this study: The baseline estimate, the residual variance ($\sigma^2$), or quality of the dataset, the duration of progress monitoring (measured in weeks), and whether or not the first observation systematically underestimated performance.

**Baseline estimate.** Five different baseline estimates were compared. Three of the baseline estimates were calculated using one day (three observations) and two were calculated using three days (nine observations) of data collection. The procedures used to estimate baseline from one day of data collection were (a) mean, (b) median, and (c) the first observation. Procedures that used three days included (a) the mean and (b) median value of the nine observations.

**Quality of dataset.** The magnitude of residual variance of WRCM estimates from a computed trend line is a way to operationally define the quality of a progress monitoring dataset. Four levels, very good ($\sigma^2 = 5$, $\epsilon^2 = 25$), good ($\sigma^2 = 10$, $\epsilon^2 = 100$), poor ($\sigma^2 = 15$, $\epsilon^2 = 225$), and very poor ($\sigma^2 = 20$, $\epsilon^2 = 400$), were used in previous simulation studies (Christ et al., 2012, 2013) and were determined through reported levels in other studies (Ardoin & Christ, 2009; Francis et al., 2008; Good & Shinn, 1990; Jenkins et al., 2005; Shinn, Good, & Stein, 1989). Christ et al. (2012) concluded that although desirable, high quality data sets are not yet observed in practice. Conversely, low quality data sets are likely to lead to such erroneous decisions they should not be used at all. To simplify the current study, only two levels of residual variance were compared. The two levels corresponded to values likely to be observed in practice; $\sigma^2 = 10$ “good” and $\sigma^2 = 15$ “poor” quality datasets.

**Duration of progress monitoring.** The number of weeks of progress monitoring was set to one of eight levels (6, 8, 10, 12, 14, 16, 18, and 20). These durations were selected because they approximated what is used in typical school-based practice and have a basis in the best practices literature (Shinn, 2008).

In total, 16 conditions were simulated (two levels of passage quality and eight levels of duration). Three hundred batches of 30 respon-
dents were generated for each of the 16 conditions with daily estimates of growth from the LMER model presented in equation 1. Data generation resulted in 9,000 total examinees per condition (144,000 total simulated cases).

**Systematic underestimation.** After the initial generation of the simulated data and subsequent model checking, and the first series of analyses, two levels of systematic bias were introduced to the first observation on the first day. The resulting biased observation was used as a baseline estimate. The first measure of oral reading fluency at the initiation of progress monitoring may not be an accurate estimate of student performance. In fact, unfamiliarity with a task is a documented source of underestimating overall performance in CBM of computational math skills (Christ & Schanding, 2007). The original model for data generation assumes that every observed WRCM is a valid indication of a student’s ability, which may not be true (Jenkins et al., 2009). Therefore, after the initial analyses, 10 and 15 WRCM were subtracted from the first observation of the first day for each simulated participant and the level of reliability, validity, bias, and precision, for the same 16 conditions was computed using only the first, biased observation, as a baseline estimate.

### Dependent Variables

Four dependent variables were computed to evaluate the reliability, validity, bias, and precision of observed growth estimates. The dependent variables were as follows: (1) the split half reliability of observed slopes, (2) the correlation between true and observed slopes, (3) mean error (ME), and (4) root mean square error (RMSE) between true and observed slopes.

**Reliability.** To determine the reliability of observed slopes, the Spearman Brown corrected (National Center for Response to Intervention, 2011; Nunnally, 1970) split half reliability was computed for each simulated respondent. For each individual, a slope was estimated using only even time points. Next, a slope was estimated using only odd time points. The two slopes were correlated to yield an estimate of reliability (Nunnally, 1970). This procedure of estimating reliability is recommended by the National Center for Response to Intervention (2011).

**Validity.** The true slope of the \( i \)th simulated subject was defined as \( \beta^*_i = \beta_0 + \beta_1i \). Estimated slope \( \beta^*_{i1} \) was obtained with OLS regression after a random residual term, corresponding to the residual variance, or the quality of the dataset, was added to each observation. The correlation coefficient between \( \beta_{i1} \) and \( \beta^*_{i1} \) provided a measure of the rank order consistency between observed and true slopes for each condition and was an index of validity.

**Bias and precision.** To estimate the precision of observed slopes, RMSE values were computed between observed and true slopes. To determine the level of bias (the systematic over or under estimation of observed slope in relation to true slope) ME values were computed. For both estimates, values that differ significantly from zero are indicative of unreliability of observed growth in indexing true growth. RMSE is a common statistic in simulation studies, and provides an indication of the magnitude of difference between predicted and true values. Mean error indicates the direction of differences. A positive ME value is indicative that observed slope estimates systematically overestimate true slope. Negative values indicate observed slope estimates systematically underestimate true slope.

### Results

#### Model Checking

The validity of the simulation methods were checked by verifying that simulated estimates of growth approximated 1.50 WRCM (the mean value specified in the data generation model) per week across conditions. Ideally, summary statistics for the simulated data should approximate the parameters specified in the model used for data generation. If parameters specified in the model are observed, this would indicate that errors did not occur when data were generated. Model checking was performed on the simulated dataset before performance of the first observation on the first day was systematically manipulated. The mean slope value across conditions was within ±.04 of 1.50 WRCM. Table 2 presents the means of the standard deviations for observed estimates of growth across conditions. The \( SD \) for true growth in the LMER
The simulation model was set to 0.63. There was a higher degree of variability of individual slopes from the group average for conditions with fewer weeks of observations and lower quality datasets. Controlling for the duration of progress monitoring and the quality of datasets, the method of baseline estimation had a small effect on the variability of $SD$s for observed estimates of growth.

**Descriptive Statistics**

Across conditions, mean and median baseline estimates when the first observation was not systematically manipulated yielded highly similar results. As a result, mean estimates are not reported.

**Reliability.** A review of the data in Tables 3 and 4 indicated that the level of reliability of growth estimates increased as the duration of progress monitoring increased. In addition, reliability estimates decreased as the level of residual variance ($\varepsilon$) of datasets increased. Baseline estimation method, even when the first observation systematically underestimated performance by 10 or 15 WRCM did not affect the variability of reliability estimates. That is reliability estimates between different baseline es-

<table>
<thead>
<tr>
<th>Weeks</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.22</td>
<td>1.55</td>
<td>1.19</td>
<td>1.00</td>
<td>0.87</td>
<td>0.81</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td>First</td>
<td>2.07</td>
<td>1.47</td>
<td>1.15</td>
<td>0.98</td>
<td>0.86</td>
<td>0.81</td>
<td>0.75</td>
<td>0.72</td>
</tr>
<tr>
<td>Good</td>
<td>3.23</td>
<td>2.20</td>
<td>1.65</td>
<td>1.35</td>
<td>1.11</td>
<td>0.99</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>3.04</td>
<td>2.11</td>
<td>1.59</td>
<td>1.31</td>
<td>1.09</td>
<td>0.97</td>
<td>0.90</td>
<td>0.84</td>
</tr>
<tr>
<td>quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2**

**Descriptive Statistics: Mean Standard Deviations for Observed Estimates of Growth**

<table>
<thead>
<tr>
<th>Weeks</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.26</td>
<td>1.57</td>
<td>1.20</td>
<td>1.01</td>
<td>0.88</td>
<td>0.80</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td>First</td>
<td>2.10</td>
<td>1.48</td>
<td>1.16</td>
<td>0.98</td>
<td>0.86</td>
<td>0.81</td>
<td>0.75</td>
<td>0.72</td>
</tr>
<tr>
<td>Good</td>
<td>3.29</td>
<td>2.24</td>
<td>1.67</td>
<td>1.36</td>
<td>1.12</td>
<td>1.00</td>
<td>0.92</td>
<td>0.85</td>
</tr>
<tr>
<td>quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>3.04</td>
<td>2.11</td>
<td>1.59</td>
<td>1.31</td>
<td>1.09</td>
<td>0.97</td>
<td>0.90</td>
<td>0.84</td>
</tr>
<tr>
<td>quality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3**

**Descriptive Statistics for Observed Estimates of Growth When Using the Median of One and Three Days as a Baseline Estimate**

<table>
<thead>
<tr>
<th>Weeks</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.24</td>
<td>.26</td>
<td>.31</td>
<td>.37</td>
<td>.47</td>
<td>.57</td>
<td>.65</td>
<td>.72</td>
</tr>
<tr>
<td>Good</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quality</td>
<td>.24</td>
<td>.26</td>
<td>.32</td>
<td>.39</td>
<td>.49</td>
<td>.58</td>
<td>.66</td>
<td>.73</td>
</tr>
<tr>
<td>Poor</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
<td>.30</td>
<td>.33</td>
<td>.38</td>
<td>.48</td>
<td>.56</td>
</tr>
<tr>
<td>quality</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
<td>.30</td>
<td>.33</td>
<td>.38</td>
<td>.48</td>
<td>.56</td>
</tr>
</tbody>
</table>

**Note.** Good and poor quality represent datasets with $\varepsilon = 10$ and 15, respectively. To estimate baseline three observations were generated per day for three days. $SE$ values were > 0.04 across all cells. RMSE-Root Mean Square Error.
timation procedures were highly similar (within two standard errors of each other).

Validity. The correlation coefficients between observed and true growth represent the rank order consistency for observed estimates and true growth. A review of the data in Tables 3 and 4 indicated that as the number of weeks of progress monitoring increased and level of residual variance (\(\varepsilon\)) of data sets decreased, correlation coefficients approached 1.00. In addition, baseline estimation method, after controlling for the number of weeks of progress monitoring and dataset quality, even when using a single biased observation, did not have an effect on the validity of growth estimates.

Bias and precision. ME values approximated zero (0.02) for every condition when data were initially generated and are not reported. When performance was systematically underestimated, observed growth was systematically greater than true growth, particularly for datasets with a high level of residual variance, and for shorter progress monitoring schedules. A review of the data in Tables 3 and 4 indicated that precision improved as the duration of progress monitoring increased and the level of residual variance (\(\varepsilon\)) decreased. Further, when progress was monitored for a short amount of time, using the median of multiple observations improved the precision of growth estimates over and above using a single unbiased observation as baseline. A more detailed examination of Tables 3 and 4 indicated that for shorter progress monitoring schedules, using the median of nine observations led to marginally more precise estimates of growth for very short durations. The improved precision declined sharply as the duration of progress monitoring increased. When the first observation systematically underestimated performance and was used as a baseline estimate, RMSE values were considerably greater than when the median of multiple unbiased observations was used as a baseline. Even as the duration of progress monitoring increased, RMSE values when using a single biased observation never approximated the values that used the median of three or nine observations as baseline.

**Discussion**

In this study, CBM-R observations were simulated in two stages. The first was a baseline approximation, and the second was a growth estimation. The accuracy of growth estimates was highly dependent on the quality of the baseline estimates. The precision of growth estimates improved as the duration of progress monitoring increased and the level of residual variance decreased. When using a single biased observation as a baseline, RMSE values were considerably greater than when the median of multiple unbiased observations was used as a baseline. Even as the duration of progress monitoring increased, RMSE values when using a single biased observation never approximated the values that used the median of three or nine observations as baseline.

**Table 4**

| Corrected split half reliability | | 
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Good quality | Poor quality | Good quality | Poor quality | Good quality | Poor quality | Good quality | Poor quality |
| Weeks | Level of bias in first observation | 0 | -10 | -15 | 0 | -10 | -15 | 0 | -10 | -15 | 0 | -10 | -15 | 0 | -10 | -15 |
| 10 | .31 | .31 | .31 | .31 | .31 | .31 | .31 | .31 | .31 | .31 | .31 | .31 | .31 | .31 | .31 | .31 |
| 12 | .36 | .36 | .36 | .36 | .36 | .36 | .36 | .36 | .36 | .36 | .36 | .36 | .36 | .36 | .36 | .36 |
| 14 | .45 | .45 | .45 | .45 | .45 | .45 | .45 | .45 | .45 | .45 | .45 | .45 | .45 | .45 | .45 | .45 |
| 16 | .55 | .55 | .55 | .55 | .55 | .55 | .55 | .55 | .55 | .55 | .55 | .55 | .55 | .55 | .55 | .55 |
| 18 | .64 | .64 | .64 | .64 | .64 | .64 | .64 | .64 | .64 | .64 | .64 | .64 | .64 | .64 | .64 | .64 |
| 20 | .71 | .71 | .71 | .71 | .71 | .71 | .71 | .71 | .71 | .71 | .71 | .71 | .71 | .71 | .71 | .71 |

**Note.** Good and poor quality represent datasets with \(\varepsilon\) = 10 and 15, respectively. SE values were > .04 across all cells. Level of bias refers to how many WRCM were deleted from the first observation on the first day. RMSE = Root Mean Square Error.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
phase where three data points were collected per day for the first three days of week one, and the second was a progress monitoring phase where one data point was collected per week. Parameters to guide the simulation were derived from two high quality extant CBM-R datasets. Baseline was calculated using as many as nine and as few as one observation. Observed estimates of growth were compared using six different baseline estimates between good and poor quality data sets and progress monitoring schedules that ranged from 6–20 weeks. In addition, after initial analyses were conducted, an instance where the first observation on the first day of baseline data collection systematically underestimated performance was modeled. Bias was induced by subtracting either 10 or 15 WRCM from that observation. The same dependent variables were then calculated when that biased observation was used as a baseline estimate.

The purpose of this study was to evaluate the effects of baseline estimation procedures on CBM-R growth estimates. Two research questions were posed: How does baseline estimation affect the reliability, validity, and precision of estimates of growth? Further, what, if any, value is added by collecting more observations to calculate baseline? The dependent variables were compared across a number of conditions, including the duration of progress monitoring (measured in weeks), the quality of datasets (or the residual variance around a computed trend line), the method used to estimate baseline, and whether the first observation on the first day was systematically biased or not. Increasing the duration of progress monitoring improved the reliability, validity, and precision of observed estimates of growth. Increasing the quality of datasets had a similar effect. Baseline estimation marginally affected the precision but not the reliability or validity of observed estimates of growth (when the first observation of the first day did not systematically underestimate performance). The results of this study suggest that collecting an excessive number of observations to estimate baseline for CBM-R progress monitoring is not advised from both a psychometric and cost–benefit standpoint.

Reliability

Reliability estimates were substantially similar at six weeks of progress monitoring when the median of nine, the median of three, and the first (unbiased) observation was used to estimate baseline performance with good quality data sets (.24, .24, and .25, respectively; see Tables 3 and 4). Reliability improved uniformly as the duration of progress monitoring increased. After 20 weeks of progress monitoring, reliability estimates for good quality data sets were .72, .73, and .71 for the three different baseline estimation procedures, respectively. Using a .70 criterion (Kelly, 1927), collecting a high number of CBM-R observations at baseline did not decrease the number of weeks before a low stakes decision could be made. Poor quality data sets yielded the same pattern of results but estimates were uniformly lower than estimates from high quality datasets, and the criterion to make low stakes decisions was never reached. Results were highly similar when a biased single observation was used as a baseline estimate.

Validity

Correlation values between observed and true slopes were not substantially different at six weeks when using the median of nine, rather than the median of three, CBM-R observations to estimate baseline with good and poor quality datasets (.29 vs. .27 and .19 vs. .18, respectively; see Table 3). This pattern was also observed at 20 weeks (.86 vs. .86 and .74 vs. .75, respectively). Highly similar patterns of results were found when using one (biased and unbiased) observation as baseline (see Table 4). Collecting a high number of observations to estimate baseline does not appear to strengthen the relationship between true and observed CBM-R growth estimates.

Bias and Precision

There was no evidence of bias among any of the baseline procedures (ME = ± .02) when the first observation did not systematically underestimate performance. Consistent with Jenkins et al. (2009), when the first observation systematically underestimated performance, and was used as a baseline estimate, observed growth was consistently greater than true growth. RMSE values were noticeably different at six weeks when using the median of nine, the median of three, and one (unbiased) observation to estimate baseline for good and poor quality data sets (see Tables 3 and 4). Table 5 demonstrated that when using a poor quality dataset, observed growth estimates can
range from negative values to almost values three times as large as true growth for shorter progress monitoring durations. Tables 3, 4, and 5 also illustrated that the differences between using the median of nine, the median of three, and the initial unbiased observation to calculate baseline attenuated sharply as the duration of progress monitoring increased. By 20 weeks the RMSE values among baseline collection procedures for good and poor quality data sets were nearly identical. When systematic underperformance was modeled in the first observation, and that value was used as a baseline estimate, RMSE values were much higher than using median values across dataset quality conditions and progress monitoring durations. For very short term progress monitoring schedules (i.e., approximately 6 weeks), RMSE values were almost double that of the results when using the median of three observations. The discrepancy in RMSE values remained substantial even after 20 weeks of data collection were simulated.

**Implications**

Before this study, there was a limited amount of research that systematically investigated how baseline estimation affects CBM-R growth estimates. Yet, it would seem that an accurate estimate of initial performance is an important preliminary step for progress monitoring. In the context of CBM-R, inadequate baseline estimation might hinder the ability to determine whether a form of instruction or academic intervention is effective at improving reading competency (Hixson et al., 2008). The degree of improvement in the reliability, validity, and precision of growth estimates as a product of collecting more baseline data needs to be considered in relation to the additional resources and missed instructional time collecting that data necessitates (Hixson et al., 2008). The results of this study are somewhat confirmatory—in that using the median of three observations collected from one day of data collection yields a satisfactory balance between the quality of progress monitoring outcomes and the amount of resources needed to collect that data.

Christ et al. (2012) found that collecting one data point per week with a good quality data set required almost 20 weeks (or 20 observations) to make low stakes educational decisions. Christ et al. (2013) found that collecting three observations per occasion could lead to low stakes decisions being made 5 weeks sooner. When compared with those results, the results of this study indicate that collecting more baseline data did not reduce the number of weeks of data collection required to make low stakes decisions compared with collecting one data point per week. In addition, low stakes decisions could not be made with poor quality data sets regardless of baseline estimation procedure. Using a dense baseline data collection and reverting back to a sparser schedule does

---

### Table 5

**Illustration of the Relationship Between Root Mean Square Error Values, Baseline Estimation Method, Progress Monitoring Duration, Observed Growth Estimates, and True Growth When Using a Poor Quality Dataset**

<table>
<thead>
<tr>
<th>Baseline estimation method</th>
<th>First (unbiased)</th>
<th>Median of three</th>
<th>Median of nine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed growth</td>
<td>Lower bound</td>
<td>Upper bound</td>
<td>Range</td>
</tr>
<tr>
<td>Weeks</td>
<td>True growth</td>
<td>Lower bound</td>
<td>Upper bound</td>
</tr>
<tr>
<td>6</td>
<td>0.10</td>
<td>-2.54</td>
<td>4.54</td>
</tr>
<tr>
<td>8</td>
<td>0.10</td>
<td>-1.30</td>
<td>3.30</td>
</tr>
<tr>
<td>10</td>
<td>0.10</td>
<td>-.64</td>
<td>2.64</td>
</tr>
<tr>
<td>12</td>
<td>0.10</td>
<td>-.26</td>
<td>2.26</td>
</tr>
<tr>
<td>14</td>
<td>0.10</td>
<td>.02</td>
<td>1.98</td>
</tr>
<tr>
<td>16</td>
<td>0.10</td>
<td>.20</td>
<td>1.80</td>
</tr>
<tr>
<td>18</td>
<td>0.10</td>
<td>.32</td>
<td>1.68</td>
</tr>
<tr>
<td>20</td>
<td>0.10</td>
<td>.42</td>
<td>1.58</td>
</tr>
</tbody>
</table>
not yield the same levels of reliability, validity, and precision as a persistently dense schedule. In other words, time and resources may not be saved by collecting more data at baseline and reverting to a more sparse progress monitoring schedule afterward.

Jenkins et al. (2009) found that different baseline estimation procedures affect the validity of growth estimates. In this study it was found that collecting more observations at baseline leads to marginally more precise estimates of growth in good and poor quality data sets for short-term progress monitoring. This finding was accentuated when the initial observation was manipulated to systematically underestimate performance. Such a scenario may be likely when a student is completing a CBM-R assessment for the first time. This finding highlights the importance of using more than one observation to estimate baseline.

Reliability and validity did not change as a function of baseline estimation procedure even when the first observation systematically underestimated performance. The longer data are collected, the less the benefits of collecting more observations at baseline are apparent. Thus, collecting large amounts of data across multiple days may not be cost-effective for schools. The minimal improvement in precision collecting such a large amount of data is arguably obviated by the resources and missed instructional time required to collect that data. The consideration is especially relevant as one of the most common criticisms of determining student responsiveness to intervention via progress monitoring is that high internal validity, applying the results of this study without considering factors of the individual student being monitored is at odds with the idiographic hypothesis testing framework inherent to CBM-R progress monitoring. For instance, if an educator is certain that a student was having an “off day,” or was not putting forth a high level of effort during the first day of data collection, the results of this study do not justify not collecting baseline data the next day. Field based data collections, including small n and single-case studies should be carried out to confirm and establish the external validity of the findings of this study.

In addition, some parameters used for data generation were not derived from the extant dataset. The quality of datasets was determined by expert judgment. As mentioned previously, the quality of a dataset is both a product of the instrumentation of the passage sets and the conditions of administration. As a result, developing recommendations for practitioners on how to choose a passage set that is likely to yield a good versus a poor quality dataset is challenging. Assuming that data are collected with a high level of integrity, it is likely that high-quality passage sets such as AIMSweb will result in high-quality datasets, and low-quality passage sets such as the original DIBELS will result in low-quality data sets (Christ et al., 2012). Identifying more of the factors that contribute to the magnitude of error, or quality of datasets requires further research.

Limitations and Future Directions

Using simulation methods enabled us to control the quality of datasets and allowed us to analyze a large amount of data that would have been difficult to collect in practice. But the results of all simulation studies and the validity of findings and implications are contingent on the legitimacy of the extant database analyzed as well as the model and parameters used to generate observations. In terms of the dataset analyzed, the results of this study are meant to generalize to predominately English speaking second and third grade students receiving Tier II interventions. Future studies should use different high quality CBM-R datasets to see whether similar results are found. Doing so would expand the generalizability of this study and other CBM-R simulation research.

Related to the generalizability of this study is the nature of the units of analysis. CBM-R growth estimates are used to make educational decisions for individual students. Each value in Tables 3 and 4 was based on 9,000 simulated respondents. The results of this study guide data collection procedures at the group level. Although the findings of this study have high internal validity, applying the results of this study without considering factors of the individual student being monitored is at odds with the idiographic hypothesis testing framework inherent to CBM-R progress monitoring. For instance, if an educator is certain that a student was having an “off day,” or was not putting forth a high level of effort during the first day of data collection, the results of this study do not justify not collecting baseline data the next day. Field based data collections, including small n and single-case studies should be carried out to confirm and establish the external validity of the findings of this study.

In addition, some parameters used for data generation were not derived from the extant dataset. The quality of datasets was determined by expert judgment. As mentioned previously, the quality of a dataset is both a product of the instrumentation of the passage sets and the conditions of administration. As a result, developing recommendations for practitioners on how to choose a passage set that is likely to yield a good versus a poor quality dataset is challenging. Assuming that data are collected with a high level of integrity, it is likely that high-quality passage sets such as AIMSweb will result in high-quality datasets, and low-quality passage sets such as the original DIBELS will result in low-quality data sets (Christ et al., 2012). Identifying more of the factors that contribute to the magnitude of error, or quality of datasets requires further research.
Finally, this study evaluated methods to estimate trend. The effect of baseline estimation on the computation of a goal line was not considered. Point decision rules require that the administrator draw a line of predicted performance from a baseline value. One method of calculating a goal line is to assume a constant rate of growth (e.g., 1.50 WRCM improvement per week). Progress is then determined if a certain number of sequential observations fall above or below that line. If a practitioner uses a constant rate of growth when creating a goal line, a higher or lower baseline observation does not affect the slope of the line, just its distance from the x axis. If the goal line is based on a large baseline estimate, observations will have to be consistently higher to make adequate progress than if the goal line was computed with a smaller baseline estimate. Although collecting more data to estimate baseline did not appear to have a substantial impact on estimates of trend, the same may not be true for decisions based on performance as measured by point decision rules in relation to a goal line. This consideration is especially relevant as point decision rules are quite common in practice (Ardoin et al., 2013).

**Summary and Conclusions**

The findings of this study suggest that collecting a high number of CBM-R observations to calculate baseline for CBM-R slope estimation may not be the best use of educators and school psychologist's time. The duration of progress monitoring and the quality of datasets are more influential than baseline estimation procedure in improving the reliability, validity, and precision of growth estimates. Collecting more data at baseline marginally increased the precision of CBM-R progress monitoring for very short progress monitoring schedules, but two additional days of data collection, especially when a large number of students are being monitored, may not be a good use of resources.

Jenkins et al. (2009) noted for CBM-R to be used with fidelity by educators, data need to be collected often enough to detect growth, but not so frequently that its use is perceived as cumbersome and detracting from instructional time. Based on the results of this study, collecting more than one day’s worth of observations to estimate baseline does not decrease the amount of time data has to be collected to yield more precise estimates of growth. The results of this study highlight the importance of improving other facets of progress monitoring practices. That is, field studies that make use of many resources (students’ time, grant money, etc.) may be better spent investigating other facets of progress monitoring. Such things as improving psychometric properties of instruments, clarifying the number of probes that ought to be administered per occasion, or how to optimally apply decision rules may be more worthwhile endeavors. We hypothesized that improving baseline estimation may be an economical and efficient way to improve the outcomes of progress monitoring. Rather, this study highlights the need to identify other facets of progress monitoring that can improve outcomes in a cost-efficient manner. Doing so will ultimately improve the defensibility of using CBM-R in a problem solving model to make high and low stakes educational decisions.

**References**


School Psychology, 47, 55–75. doi:10.1016/j.jsp 2008.09.004


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, 5–27. doi:10.1300/ J008v02n02_02


Jenkins, J., & Terjeson, K. J. (2011). Monitoring reading growth: Goal setting, measurement frequency, and methods of evaluation. Learning Dis-
abilities Research & Practice, 26, 28–35. doi:10.1111/j.1540-5826.2010.00322.x


Received November 12, 2012
Revision received March 2, 2013
Accepted March 6, 2013

E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at http://notify.apa.org/ and you will be notified by e-mail when issues of interest to you become available!