School Psychology Review,
2016, Volume 45, No. 3, pp. 296–309

Curriculum-Based Measurement of Reading: Accuracy of Recommendations From Three-Point Decision Rules

Ethan R. Van Norman
Georgia State University

Theodore J. Christ
University of Minnesota

Abstract. Despite their widespread use, there is little research to support the accuracy of curriculum-based measurement of reading progress monitoring decision rules. The purpose of this study was to investigate the accuracy of a common data point decision rule. This study used a three-point rule with a goal line of 1.50 words read correctly per minute (WRCM) across six levels of true growth (range = 0–3 WRCM), two levels of dataset quality or residual (5 and 10 WRCM), and 13 levels of data collection (range = 3–15 weeks). We estimated the probability of a correct decision as well as the probability of each outcome (change instruction, increase the goal, maintain instruction) across each condition with probability theory and a spreadsheet program. In general, results indicate that recommendations are often inaccurate. Further, the probability of a correct recommendation is below chance in most situations. Results of multiple regression analyses indicate that residual, duration, and true growth interacted to influence decision accuracy. Results are discussed along with implications for future research and practice.

Curriculum-based measurement (CBM) is used to index the level and rate of academic performance in the basic skill areas of reading, mathematics, written expression, and spelling (Deno, 1985). There are a variety of educational measures that are useful to index the level of performance, but CBM is often described as a procedure that is uniquely useful to monitor individual student progress and evaluate instructional programs (Deno, 1986). That unique utility emerged because the procedures were intentionally developed to be

Theodore J. Christ, PhD, has equity and royalty interests in, and will serve on the Board of Directors for, FastBridge Learning (FBL), 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. These interests have been reviewed and managed by the University of Minnesota in accordance with its conflict-of-interest policies.

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

Correspondence concerning this article should be sent to Ethan R. Van Norman, Department of Counseling and Psychological Services, Georgia State University, Atlanta, GA 30302; e-mail: evannorman@gsu.edu

easy to administer, efficient, technically adequate, repeatable, and useful to teachers (Deno, 2003). As a result, CBM enables teachers and other educators to collect data across time, plot the results on a time-series graph, and estimate the general trajectory of student achievement (Deno, 1986, 1990). The interpretation of those data are intended to guide instructional decisions. Such decisions often include continuing, modifying, or terminating instructional programs (Deno, 1986, 1990, 2003).

Progress monitoring is the evaluation of instructional effects using time-series data, which was the original intended application of CBM (Deno, 1985, 2003). One early and frequently cited meta-analysis estimated that the mean effect size for progress monitoring was 0.70 (Fuchs & Fuchs, 1986), which is often described as a medium to large effect. As part of that meta-analysis, the authors identified four critical components connected with the effect: (a) ongoing data collection, (b) graphic displays of observations, (c) explicit decision rules to determine when instructional programs ought to be changed, and (d) behaviorally based interventions. The results of that review do not confer support for CBM specifically; rather, they confer support for progress monitoring generally. In the discussion, the authors of the study emphasized that graphical displays and decision rules were essential for effective progress monitoring: "When teachers were required to employ data utilization rules, effect sizes were higher than when data were evaluated by teacher judgment" (p. 205).

In contemporary research, data utilization rules are described as decision rules. Two subsequent reviews supported the conclusion that explicit decision rules are necessary to facilitate the interpretation and use of progress monitoring data (Stecker & Fuchs, 2000; Stecker, Fuchs, & Fuchs, 2005). Generally speaking, *interpretation* is the act of deriving meaning and *use* is the act of applying that interpretation to an educational decision (Kane, 2013). Teachers and school psychologists often use progress monitoring data to make routine low-stakes easily reversible educational decisions. Examples of low-stakes decisions include modifying, maintaining, or terminating an instructional program. The original intent of CBM was to guide such low-stakes decisions for individual student special education programming (Deno, 1986). More recently, educators and school psychologists have been using progress monitoring data to make higher stakes decisions or decisions that are not as easily reversible. Examples of such decisions may be tier placement or diagnostic decisions within a response-to-intervention model of special education eligibility determination (Vaughn & Fuchs, 2003). At the time of this study, federal law allowed for the use of progress monitoring data to inform special education eligibility decisions (Individuals With Disabilities Education Improvement Act, 2004).

DECISION RULES

Evidence to support and evaluate CBM of oral reading (CBM-R) decision rules was summarized and evaluated in a recent review (Ardoin, Christ, Morena, Cormier, & Klingbeil, 2013). A literature search identified 102 published documents that met inclusion criteria for the review. As an important note, the review excluded studies that examined instructional effects or student outcomes if there were no analyses of the technical qualities of data and decision rules (e.g., Fuchs & Fuchs, 1986; Stecker & Fuchs, 2000; Stecker et al., 2005). The focus of the review was specific to the technical adequacy of CBM-R time-series data and decision rules. The review identified two categories of decision rules: data point and trend line.

The data point rule begins with a goal line, or an expected rate of weekly improvement. At any point in time, CBM-R observations above the goal line are desirable and observations below the goal line are undesirable. One type of data point rule indicates that (a) if three consecutive data points fall below the goal line, the current instructional strategy should change; (b) if three consecutive data points fall above the goal line, the goal should be increased; and (c) if three consecutive data points are distributed around the goal line, the

instructional strategy should continue. Common variations on the data point rule are to use the most recent four or five data points. As an alternative, the trend line rule compares the trajectory, or slope, of the goal line with the slope of a trend line fitted through all observations within a phase, or instructional condition. In current practice, the rates of improvement (ROIs) for the goal line and the trend line are each quantified as words read correctly per minute (WRCM) gained per week. For example, the goal line might establish an expectation of 1.50 WRCM per week. It follows that a trend line of 1.25 WRCM per week is below the goal and 1.75 WRCM per week is above the goal.Ardoin et al. (2013) "did not identify any study [up through 2010] that evaluated the accuracy of the data point or trend line decision rules as related to CBM-R progress monitoring data" (p. 12). Notwithstanding, decision rules are frequently described and recommended within the professional literature. The review identified 59 published documents that described decision rules; data point and trend line rules were the most common (Ardoin et al., 2013). The most frequently cited source for data point decision rules was a non-empirical book chapter (White & Haring, 1980) and a published study that evaluated student outcomes, not the reliability or validity of decisions (Fuchs, Fuchs, & Hamlett, 1989). White later informed the second author of this study that Katherine Liberty developed the data point decision rule in 1972 for her dissertation, but very little research was done to validate its use (O. R. White, personal communication, February 2, 2011). The two most frequently cited sources for trend line decision rules also did not evaluate the reliability or validity of decisions (Good & Shinn, 1990; Shinn, Good, & Stein, 1989).

The results of the review by Ardoin et al. (2013) illustrated the need for research to evaluate the reliability and validity of decisions. Shortly after that review, researchers conducted simulation studies that evaluated the reliability and validity of trend line rules (Christ, Zopluoglu, Long, & Monaghen, 2012; Christ, Zopluoglu, Monaghen, & Van Norman, 2013). One of the key findings from those studies was that unwanted variability in student performance, or error, negatively affected the reliability and validity of trend line rules. This study extends work on decision rules by examining the data point decision rule.

VARIABILITY IN PERFORMANCE

CBM-R was developed to be a highly sensitive measure of student performance (Deno, 1986, 2003). As such, individual student data are often highly variable across repeated administrations. If CBM-R were a perfect index of instructional effects, the quality (or lack thereof) of instruction would be the only source of variability in observations across time (Poncy, Skinner, & Axtell, 2005). However, CBM-R is sensitive to factors irrelevant to instruction. Several sources of unwanted variability can be attributed to factors practitioners and researchers can control such as instrumentation (Francis et al., 2008), administration setting (Derr-Minneci & Shapiro, 1992), and administration directions (Colon & Kranzler, 2006). In addition, CBM-R is likely to be sensitive to factors beyond researchers and practitioners' control such as variations in a student's motivation, disposition, and alertness. Attaining consistent and comparable performances across repeated administrations is difficult to accomplish. It requires high-quality instrumentation composed of alternate forms of equivalent difficulty along with tightly standardized administration conditions.

Unexplained variation in performance is conceptualized as error, or residual. The standard error of measurement for CBM-R often approximates 6 to 12 WRCM (Christ & Ardoin, 2009; Christ & Silberglitt, 2007; Poncy et al., 2005). That value is useful to construct a confidence interval around a single score. It also happens that the variation in student performance around the trend line often approximates 6 to 12 WRCM (Ardoin & Christ, 2009; Christ, 2006; Hintze & Christ, 2004). That variation around the trend line is referred to as the *standard error of the estimate* (*SEE*). The *SEE* is assumed to be normally distributed. With an *SEE* of 10 WRCM, 68% of the

data at any given time point would fall ± 10 WRCM of the trend line, or within a trended envelop of 20 WRCM across the entire time series. Recent studies examined the influence of such variation on the reliability and validity of trend line decision rules using simulation methodology (Christ et al., 2012, 2013). Similar studies are necessary to examine the reliability and validity of the data point decision rule.

PURPOSE

The purpose of the current study was to examine the accuracy of data point decision rules to help establish evidence-based guidelines for their use. Given that progress monitoring outcomes inform high-stakes decisions such as special education eligibility, incorrect decisions have numerous potential negative consequences. For instance, if a decision rule suggests that a student is not improving at an adequate rate and an intervention is in fact effective, a successful instructional strategy may be inappropriately abandoned. Such an outcome may seem trivial, but making meaningful instructional modifications based on individual student progress requires substantial resources (Stecker & Fuchs, 2000; Stecker et al., 2005). Likewise, incorrectly identifying a student as not improving increases the chances that he or she will be misdiagnosed as having a learning disability. If a decision rule suggests that a student is improving when in fact he or she is not, ineffective instructional strategies are likely to persist. Thus, the discrepancy between a target student's performance and peers will only continue to widen, even as he or she receives (seemingly) effective supplemental supports. Furthermore, incorrectly inferring a student is making adequate progress increases the likelihood more appropriate intensive supports (i.e., special education services) will be withheld.

Recent research has suggested that trend line rules are unlikely to yield reliable and valid interpretations or accurate educational decisions, especially if data are collected over a brief period (e.g., 6 weeks) and *SEE* is large (e.g., >10 WRCM; Christ et al., 2012, 2013). Such findings warrant further investigations as to what progress monitoring practices increase the likelihood of making accurate decisions, or at the very least verify current recommendations derived from expert opinion.

It was expected that the probability of a correct decision was close to chance in the first few weeks of progress monitoring. We predicted that the probability of a correct decision would increase in relation to the magnitude of the intervention effect, the magnitude of residual, and the duration of progress monitoring. More specifically, decision accuracy would be modest when true growth differed substantially from the goal line, residual was small, and durations were long. In contrast, decision accuracy would be low when the true ROI approximated the goal line, residual was large, and duration was short. We evaluated the probability of a correct decision within six levels of true ROI (range = 0–3 WRCM increase per week), two levels of residual (5 and 10 WRCM), and 13 durations (range = 3–15 weeks).

METHOD

We derived the probability that the three-point decision rule would result in a correct decision across a large number of progress monitoring scenarios. We based those conditions (described in the Design subsection) on a previous analysis of a large extant progress monitoring dataset.

Participants

The dataset consisted of 1,517 second-grade and 1,561 third-grade students. The demographic makeup of the sample was as follows across grades: 46% girls and 53% White, 17% Black, 8% Hispanic or Latino, 6% Asian or Pacific Islander, and 2% American Indian or Alaska Native. Approximately 2% of participants within each grade received special education services.

Procedure—Extant Dataset

The extant dataset was obtained via an agreement between the second author and a

state coordinator of a federally funded program that provided supplemental (Tier 2) standard-protocol, evidence-based reading interventions to elementary students with reading difficulties. Students were identified for the program if they scored below a predetermined CBM-R benchmark as part of school-wide universal screening. Local schools attained parental consent in coordination with the agency.

As part of the program, a hired data collector administered one grade-level AIMSweb probe per week to monitor the effects of the intervention. Specific information about the nature and intensity of the interventions, as well as the data utilization rules used at the time, was unavailable to the authors. Data collectors were hired by the state agency and were trained to criterion with AIMSweb training materials and assessed for administration fidelity using the Accuracy of Implementation Rating Scale (Shinn & Shinn, 2002). Specific administration fidelity data were not available, but acceptable scores (95% or greater) on the scale were a condition for continued employment. Interrater reliability data were also not available, but published estimates typically approximate or exceed 0.95 (Wayman, Wallace, Wiley, Ticha, & Espin, 2007). Data were deidentified at the school, student, and administrator level prior to analysis.

Design

We used a 6 × 2 × 13 fully crossed factorial design, with six levels of true ROI (range = 0–3 WRCM per week), two levels of residual (5 and 10 WRCM), and 13 durations (range = 3–15 weeks). The distribution of true ROI values was selected based on the results of a linear mixed effects regression (LMER) model estimated to the extant progress monitoring dataset. One method of specifying a goal line for decision making is to use average growth rates from normative tables. The slope of the goal line in this study was set to the average ROI of participants from the LMER analysis, which was 1.50 WRCM per week. The goal ROI also corresponds with the typical value observed for students in evidence-based instructional programs that are implemented with high fidelity (Deno, Fuchs, Marston, & Shin, 2001; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993).

True Growth

Previous work defined true and observed growth as it relates to CBM-R progress monitoring data (Christ et al., 2012, 2013; Jenkins, Graff, & Miglioretti, 2009). In this study the true ROI was the value that would be observed if there were no measurement error. For the purpose of this study, true growth was defined at six levels: 0.00, 0.84, 1.25, 1.75, 2.16, and 3.00 WRCM per week. Those ROIs corresponded with 1st, 15th, 30th, 70th, 85th, and 99th percentile values in the extant dataset, respectively.

Residual

Residual was described in the introduction. For the purpose of this study, it was set to one of two levels: 5 and 10 WRCM. These values were selected because they are generally used to describe very good-quality datasets and good-quality datasets, respectively, in the research literature (Ardoin & Christ, 2009; Christ, 2006; Christ et al., 2012, 2013; Hintze & Christ, 2004). Residual values are analogous to *SEE* values and are indicative of the typical variability of observations within a progress monitoring case. This is akin to measurement error, which obscures estimates of true growth.

Duration

Duration was the number of weeks of progress monitoring before a decision was made. The length of progress monitoring was set to 1 of 13 levels. The shortest duration was 3 weeks because three data points are required to apply the decision rule. The longest duration was 15 weeks.

Correct Decision

The goal line slope was set to 1.50 WRCM per week. The correct decision was to change instruction when the true ROI was <1.50 WRCM per week, increase the goal if the true ROI was >1.50 WRCM per week, or maintain instruction if the true ROI was equal

to the goal ROI. As described earlier, the true ROI was specified for each case. As a result, the correct decision was always known.

Analyses

We estimated the probability of each outcome (change instruction, increase the goal, maintain instruction) and compared the resulting recommendation with the correct decision. All calculations were run in a Microsoft Excel spreadsheet that is available from the first author. The analytic procedure was a derivation not a simulation. That is, neither progress monitoring cases nor CBM-R scores were generated for the analysis. Instead, we used probability theory to derive the likelihood that three consecutive observations would fall below or above the goal line given a common intercept, specified levels of true growth, residual, and duration. Specific details on how we estimated the probability that each observation would fall above or below the goal line, as well as how we estimated the cumulative probability to evaluate the recommendation from the data point decision rule, are described in the following paragraphs.

True performance at each week was calculated as the product of week number and true ROI. Assuming an intercept of 40 WRCM and true ROI of 0.84 WRCM per week, true performance at 10 weeks was 48.40 WRCM:

True: 48.80 WRCM = 40 WRCM + 0.84 WRCM per week × 10 weeks

By using the same intercept, the expected performance based on the goal line was 55:

Goal: 55 WRCM = 40 WRCM + 1.50 WRCM per week × 10 weeks

The true performance is less than the goal performance in this example.

Residual was set to one of two values (5 or 10 WRCM). Residuals were assumed to be uncorrelated across time and normally distributed (or centered) around true performance at each week. The standard deviation of the distribution was equal to the residual (*SD* = residual within each condition; i.e., 5 or 10 WRCM). Such assumptions are consistent with the application of ordinary least squares regression, as well as many statistical procedures (Cohen & Cohen, 1983). With that, the probability that CBM-R performances would fall above or below the goal line was derived (not simulated) each week by calculating the area of the distribution that fell above the goal line and below the goal line for that observation. The product of each set of three consecutive probabilities (three CBM-R below goal, three CBM-R above goal, or at least one CBM-R above and below) provided the probability of a correct or incorrect decision. For instance, at Week 6, the probabilities (p) at Weeks 4, 5, and 6 were used to estimate the probability of an instructional change, $p_{I;}$ probability of a goal change, p_G; or probability to maintain, p_M. The calculations were

$p_I = p$ Week 4 was below × p Week 5 was below × p Week 6 was below

$p_G = p$ Week 4 was above × p Week 5 was above × p Week 6 was above

$p_M = 1 - PI + PG$

The probability of an accurate decision was derived for each of 158 unique conditions in the factorial design (Table 1).

Subsequently, multiple regression was used to estimate the amount of unique variation of decision accuracy that was associated with each independent variable: true ROI, residual, and duration (Table 2). For the purpose of these analyses, true ROIs were coded as the difference between the goal ROI and true ROI. For example, if the goal ROI was 1.50 and true ROI was 0.84, then the true ROI difference was 0.66 (1.50 – 0.84 WRCM per week). Duration was centered at 3 weeks.

RESULTS

Visual inspection of Table 1 indicated several patterns. First, a main effect for residual was apparent. Across all levels of true ROI and duration, the probability of a correct decision decreased when residual was 10 WRCM. For instance, when progress was monitored for 6 weeks and the true ROI was equal to 0.84 or 2.16 WRCM per week, the probability of a correct decision was .35 when residual was equal to 5 WRCM and .22 when residual was equal to 10 WRCM. Regression

Table 1. Probability of Correct Decision

True ROI (Percentile), WRCM per Week	Residual, WRCM	Week 3	4	5	6	7	8	9	10	11	12	13	14	15
		Probability of Correct Decision to Change Instruction												
0.00 (1st)	5	**.23**	**.37**	.53	.67	.79	.87	.93	.97	.99	1.00	1.00	1.00	1.00
	10	**.17**	**.23**	**.30**	**.38**	**.46**	.54	.61	.68	.74	.80	.85	.88	.91
0.84 (15th)	5	**.17**	**.22**	**.28**	**.35**	**.41**	**.49**	.56	.62	.69	.74	.80	.84	.88
	10	**.15**	**.17**	**.19**	**.22**	**.25**	**.28**	**.31**	**.34**	**.38**	**.42**	**.46**	**.49**	.52
1.25 (30th)	5	**.14**	**.16**	**.18**	**.19**	**.22**	**.24**	**.26**	**.28**	**.31**	**.33**	**.36**	**.38**	**.41**
	10	**.13**	**.14**	**.15**	**.16**	**.17**	**.18**	**.19**	**.20**	**.21**	**.22**	**.23**	**.24**	**.25**
		Probability of Correct Decision to Increase Goal												
1.75 (70th)	10	**.13**	**.14**	**.15**	**.16**	**.17**	**.18**	**.19**	**.20**	**.21**	**.22**	**.23**	**.24**	**.25**
	5	**.14**	**.16**	**.18**	**.19**	**.22**	**.24**	**.26**	**.28**	**.31**	**.33**	**.36**	**.38**	**.41**
2.16 (85th)	10	**.15**	**.17**	**.19**	**.22**	**.25**	**.28**	**.31**	**.34**	**.38**	**.42**	**.46**	**.49**	.52
	5	**.17**	**.22**	**.28**	**.35**	**.41**	**.49**	.56	.62	.69	.74	.80	.84	.88
3.00 (99th)	10	**.17**	**.23**	**.30**	**.38**	**.46**	.54	.61	.68	.74	.80	.85	.88	.91
	5	**.23**	**.37**	.53	.67	.79	.87	.93	.97	.99	1.00	1.00	1.00	1.00

Note. The table shows the probability of a correct decision using a three-data point decision rule conditioned on true rate of improvement (ROI), residual, and duration. Boldface values indicate conditions where the probability of a correct decision was less than chance (.50). For all analyses, we used a data collection schedule where one observation was collected per week and compared with a 1.50 words read correctly per minute (WRCM) goal line using a three-point decision rule.

analysis suggested that modeling residual led to a statistically significant improvement in model fit relative to the null model, $F(1, 154) = 104.69$, $p < .001$. Relatedly, residual accounted for approximately 9% of the variability in the probability of a correct decision (see Table 2; $R^2 = .09$).

Second, there appeared to be a main effect for duration (see Table 1). That is, across true ROI magnitudes and residual levels, as the duration of progress monitoring increased, the probability of a correct decision also increased. For instance, when the true ROI was equal to 0.00 or 3.00 WRCM per week and residual was equal to 5 WRCM, the probability of a correct decision at Week 5 was equal to .53. Within the same conditions, at 8 weeks, the probability of a correct decision jumped to .87. At 11 weeks, the probability of a correct decision jumped further to .99. Adding duration as a predictor in the multiple regression analysis resulted in a sharp increase in the explained variance of the probability of a correct decision ($R^2 = .41$; see Table 2) and a statistically significant improvement in model fit, $F(1, 153) = 357.88$, $p < .001$. In other words, duration accounted for approximately 32% of the unexplained variability in the probability of a correct decision not explained by residual.

Third, true ROI influenced the probability of a correct decision. When the true ROI approximated the slope of the goal line (1.25 and 1.75 WRCM per week), the probability of a correct decision never exceeded chance levels, regardless of residual level or the duration of data collection. Conversely, when the ROI was equal to 0 or 3 WRCM per week, the probability of a correct decision exceeded chance levels after only 5 weeks of data collection when residual was 5 WRCM and 8 weeks when residual was 10 WRCM. The

Table 2. Predicting Probability of Correct Decision

	Null Model		Model 1		Model 2		Model 3		Final Model	
Predictor	B	*SE*	B	*SE*	B	*SE*	B	*SE*	B	*SE*
Intercept	.450**	.020	.530**	.030	.280**	.040	.000	.780	.040	.030
Main effect										
R			–.030**	.001	–.030	.001	–.030**	.003	.010**	.010
D					.040**	.004	.040**	.002	.030**	.003
True ROI							.360**	.020	.250**	.030
Interactions										
R × D									–.010**	.001
R × ROI									–.040**	.008
D × ROI									.030**	.004
R × D × ROI									.003**	.001
Adjusted R^2				.09		.41		.86		.94

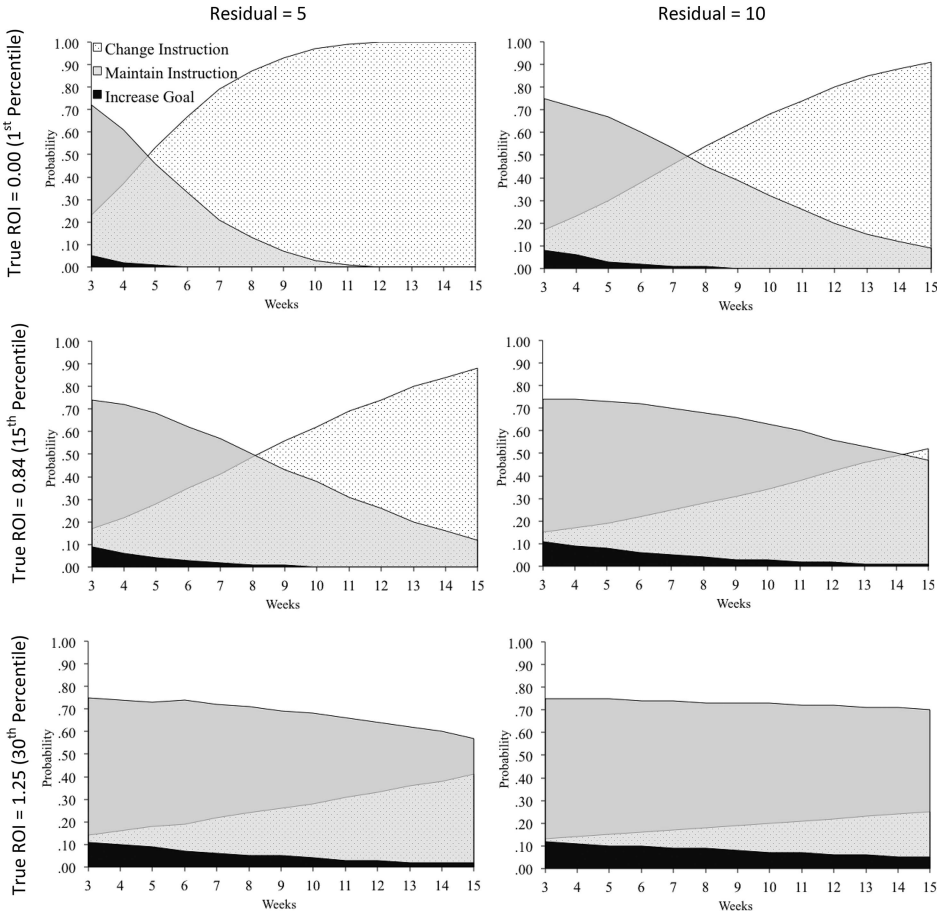
Note. The table shows predictors for the probability of a correct decision using a data point decision rule conditioned on residual (R), duration (D), and true growth. Rate-of-improvement (ROI) values were coded to reflect the absolute difference between the true ROI and the slope of the goal line (1.50 words read correctly per minute). Duration corresponded to the number of weeks of data collection, in which one observation was collected per week.
**$p < .01$.

regression analysis also suggested that as the absolute difference of ROI from the slope of the goal line increased, the probability of a correct decision increased. In fact, adding ROI as a predictor in the analysis resulted in a sharp increase in the explained variance of the probability of a correct decision ($R^2 = .86$; see Table 2) and a statistically significant improvement in model fit, $F(2, 152) = 500.89$, $p < .001$. That is, true ROI accounted for an additional 45% of the unexplained variability in the probability of a correct decision not explained by residual or duration.

The final multiple regression analysis included a three-way interaction between residual level, duration, and true ROI difference (see Table 2). The three-way interaction term was statistically significant. In essence, the interaction between residual level and duration of progress monitoring differed across levels of true ROI. Indeed, visual inspection of Table 1 indicated that the interaction between residual level and duration of progress monitoring depended on the magnitude of ROI. For instance, when ROI was 0 or 3 WRCM per week, the probability of a correct decision exceeded chance after 5 weeks and 8 weeks when residual was equal to 5 and 10 WRCM, respectively. When ROI was 0.84 or 2.16 WRCM per week, the probability of a correct decision exceeded chance levels after 8 weeks and 14 weeks (6 weeks longer) when residual was equal to 5 and 10 WRCM, respectively. Last, chance levels were never exceeded when ROI magnitude was 1.25 or 1.75 WRCM per week regardless of duration or residual. Modeling the three-way interaction increased the explained variance of the probability of a correct decision ($R^2 = .94$), as well as a statistically significant improvement in model fit, $F(4, 148) = 49.43$, $p < .001$. Modeling the three-way interaction between residual, duration, and true ROI accounted for an additional 13% of the unexplained variability in the probability of a correct decision not explained by modeling residual, duration, and true ROI as main effects.

Estimating the probability of specific treatment decisions may shed more light on the repercussions of each independent variable on student outcomes. For instance, it may be useful to know the probability that an ineffective intervention will be continued or the likelihood that a goal would be increased when the

Figure 1. Probabilities



Note. The figure shows the probability of changing instruction, maintaining instruction, and increasing the goal using a three–data point decision rule with a 1.50-words read correctly per minute per week goal conditioned on the true rate of improvement (ROI), residual, and duration.

student is in fact struggling and an instructional change should be made.

Figure 1 presents the probability of changing instruction, maintaining instruction, and increasing the goal for a three-point decision rule across durations conditioned on residual levels for three levels of true ROI: 0.00, 0.84, and 1.25 WRCM per week. These values reflected the 1st, 15th, and 30th growth percentiles, respectively. Across all panels of Figure 1, the correct decision was to change instruction. Visual inspection of Figure 1 indicated that across residual levels, as true ROI approximated the slope of the goal line, the probability of incorrectly maintaining instruction increased. The probability of increasing the goal never exceeded 0.10 across all levels of residual and true ROI. When true ROI approximated 1.25 WRCM per week, the probability of maintaining ineffective instruction exceeded the probability of changing instruction through 15 weeks. However, when true ROI was equal to 0.00 or 0.84 WRCM per week, a clear inverse relationship between changing instruction (the correct choice) and maintaining instruction (an incorrect choice) as a function of duration emerged. That is, there was a clear point where the probability

of correctly changing instruction overtook the probability of incorrectly maintaining instruction. The strength of the relationship weakened, or the point of reversal occurred at later durations, as residual increased. For instance, when true ROI was equal to 0.00 WRCM per week and residual was equal to 5 WRCM, the probability of changing instruction overtook the probability of maintaining instruction at about 4–5 weeks. When residual was equal to 10 WRCM, the point of reversal occurred at about 7–8 weeks. The effect of residual was more pronounced when true ROI was equal to 0.84 WRCM per week. The probability of changing instruction overtook the probability of maintaining instruction at about 7–8 weeks when residual was equal to 5 WRCM. That value increased to approximately 13–14 weeks when residual was equal to 10 WRCM. While Figure 1 only presented scenarios where the correct choice was to change instruction, the same inferences can be made for true ROIs that were >1.50. That is, the same pattern of results was observed when evaluating the probability of increasing the goal compared with the probability of incorrectly maintaining instruction.

DISCUSSION

The ability to make accurate interpretations of student progress is foundational to data-based decision making. CBM, in particular CBM-R, is one of the most commonly used assessments to monitor student progress (Wayman et al., 2007). Despite its widespread use, few investigations have explored the technical adequacy of common interpretive guidelines. The purpose of this study was to evaluate the accuracy of data point decision rules when applied to CBM-R progress monitoring data. We explored the accuracy of a three-point rule using a 1.50 WRCM per week goal line across six levels of true ROI, two levels of residual, and 13 levels of duration. The probability of each decision and whether it was correct was derived for 158 unique conditions. The three possible decisions were to change instruction, maintain instruction, or increase the goal. The true ROI was specified within each condition, so the correct decision was always known.

As the deviation of true ROI increased from the slope of the goal line, the duration of progress monitoring increased, and the level of residual decreased, the probability of a correct decision increased. Furthermore, the interaction between residual level and duration of progress monitoring differed as a function of true ROI.

Residual accounted for a significant, albeit small, proportion of unique variability in decision accuracy (9%). Although it is advisable to minimize extraneous variability in student performance across time, these efforts are likely to have only a modest contribution to improve the accuracy of decisions when using a data point rule. It seems that the selection of instrumentation, the setting of administrations, and the qualities of standardized administrations can only improve the accuracy of decisions to a small degree. It is very important to emphasize that this study derived estimates of accuracy for only very good-quality datasets and good-quality datasets with residuals of 5 and 10 WRCM.

The duration of progress monitoring is influential. The number of weeks accounted for substantially more unique variance in decision accuracy (32%), and an interaction was observed between residual level and duration. Therefore, if residual is large, data will have to be collected for longer durations to make an accurate decision. As a result, an either–or approach is not advised. It is necessary to both control for residual with good-quality instrumentation, as well as conditions, and collect data for longer durations. Short durations are rarely advisable, particularly not <12–14 weeks.

After accounting for both residual and duration, the deviation of the true ROI from the goal ROI accounted for a significant and large proportion of unique variance in decision accuracy (45%). That is, the probability of an accurate decision increases as the student's underlying ROI deviates more from the goal ROI. The probability of an accurate decision is less than chance when the true ROI approximates the goal ROI (1.50 WRCM per week).

The probability of a correct decision was >50% only if the difference between true and goal ROIs was >0.25 WRCM per week. That is, probabilities above chance were observed only when true ROI was <1.25 WRCM per week (30th percentile) or >1.75 WRCM per week (70th percentile). The data point rule functions only moderately well when intervention effects are very large or nonexistent.

The probability of each type of decision was derived (i.e., change instruction, maintain instruction, increase the goal). If the true ROI was less than the goal ROI, then the inaccurate decision to maintain instruction was more likely than the correct decision to change instruction for the first 5 to 15 weeks (Figure 1). The likelihood was a function of all three variables: true ROI, duration, and residual. Longer durations were necessary when the true ROI was less discrepant from the goal ROI and when residual was 10 rather than 5 WRCM. The incorrect decision to increase the goal was unlikely when the true ROI was less than the goal ROI. A similar pattern was observed for the inverse conditions, or when the true ROI was greater than the goal ROI. The inaccurate decision to maintain instruction was more likely than the correct decision to increase the goal for the first 5 to 15 weeks, depending on the conditions. In general, when residual is high and duration is short, an incorrect decision is likely. Moreover, the incorrect decision is often to maintain the presumably ineffective intervention.

Implications for Practice

Data point rules do not improve the reliability and validity of decisions relative to trend line rules. On the basis of the results of this study, practitioners can increase the likelihood of a correct decision in several ways. First, they can minimize residual by seeking out high-quality instruments and following administration directions. Generally, CBM-R probes created by commercial vendors will be of sufficient quality. At the very least, practitioners should not randomly select passages from grade level reading materials. Relatedly, stating the same standardized directions for each administration and conducting the assessment in a distraction-free quiet environment will likely help reduce residual. Practitioners can estimate the residual or *SEE* of a progress monitoring case using Microsoft Excel with the STEYX function selecting WRCM scores for *known y's* and the appropriately coded data collection day for *known x's*. If the resulting *SEE* is substantially greater than 5 WRCM, one should ensure that data have been collected for an appropriate duration. If the *SEE* is substantially greater than 10 WRCM, the practitioner is unlikely able to use said data to make a decision. Relatedly, practitioners should abstain from making decisions until collecting 12–14 weeks of data, especially if only one CBM-R is collected per week. However, the results of this study and studies similar to it suggest that even when residual is low, the probability of a correct decision is unlikely for extremely short data collection schedules.

Last, practitioners should use trend line decision rules until evidence for improved decision rules emerges. Although trend line decision rules are not perfect, they are in general more accurate than data point decision rules. With the advancement of computer technology, trend lines can be calculated with a few keystrokes. If practitioners are constrained to using data point decision rules, they should at least default to maintaining the current instructional program if the general pattern of observations approximates the goal line.

Implications for Research

The findings of this study have several implications for research. In combination with recent research on trend line rules, current interpretative methods for CBM-R progress monitoring data do not support decisions regarding instructional effects for individual students across relatively brief periods. The results of this study also suggest that residual is in fact influential on decision accuracy especially when using a data point rule. As a result, researchers need to continue to improve instrumentation and guidelines for data collection to minimize residual. At the moment, the

most potent way to improve the accuracy of decisions is to collect data for longer durations. However, waiting to make decisions for upwards of 3 months for a majority of students is counterintuitive to the premise of formative assessment. As a result, researchers should specifically focus on developing decision rules that allow educators to make accurate decisions in a reasonable amount of time.

This study evaluated the accuracy of data point decision rules applied to CBM-R progress monitoring data. While CBM-R is the most popular form of CBM, there is a paucity of research evaluating the accuracy of decision rules applied to other forms of CBM. Sensitivity to improvement, normative growth rates, and residual differ across CBM types. As a result, it is unclear whether different decision rules may be more appropriate for different types of CBM.

Last, the prevalence of the use of different decision rules in schools is unclear. Trend line rules are overwhelmingly recommended in the research literature (Ardoin et al., 2013), yet state departments of education still allow for the use of data point decision rules when evaluating students for special education services (e.g., Iowa Area Education Agencies, 2014). Future research should investigate the prevalence of different decision rules in schools.

Limitations

The design and analytic methods relied on the assumptions that (a) residuals were normally distributed and uncorrelated across time, (b) one CBM-R was collected each week, (c) true ROI was monotonic and linear, and (d) both the true and goal lines had the same intercept. Each of these assumptions is reasonable and consistent with prior research and practice; however, future research is necessary to examine variations on these assumptions and how they would affect the results of these findings.

In addition, study conditions were based on an analysis of an extant dataset. As a result, we did not have access to a host of information that may have affected the quality of the data we analyzed. More specifically, we did not have access to information such as interrater reliabilities, demographic information of individual students, data utilization strategies used by schools, the frequency and intensity of supplemental interventions, and which interventions were used as part of standard protocol treatments. Thus the generalizability of the current results may be limited. Within the CBM-R progress monitoring literature, the relationship between the accuracy of decisions and the type and intensity of interventions remains unclear. As a first step, future research needs to investigate the relationship between specific intervention protocols, as well as intervention intensity, and normative growth rates. It may be that different growth rates result from different interventions. Similarly, growth rates may differ as a function of intervention intensity. If such findings are observed, ubiquitous decision rules may not be appropriate for CBM progress monitoring data.

CONCLUSION

This study is another in a line of inquiry that addresses the technical adequacy of interpretations and decisions that might result from CBM-R progress monitoring. There is much work to be done to improve the state of affairs. In the interim, those who use CBM-R and progress monitoring are advised to use skilled visual analysis and professional judgment in combination with statistical analysis. Researchers have developed workshops to train school psychologists and educators to integrate visual and statistical analysis when interpreting single-subject data (e.g., Barton, Ferron, Kratochwill, Levin, & Machalicek, 2014; Williams & Hunley, 2015). In addition, several books are available that address the topic (Burns, Riley-Tillman, & Gibbons, 2013; Riley-Tillman & Burns, 2009). Within the context of CBM-R progress monitoring, special attention should be paid to estimates of intercept, slope, and standard errors. The promise of idiographic data-based approaches to improve student outcomes is substantial and a cornerstone of effective school-based service

delivery (Deno, 1990). This work and related work contribute to refine our knowledge and the underlying methodology of making accurate decisions for individual student programming. At the time of this study, there were no viable researched alternatives to traditional decision rules; however, our teams and others are working on new analytic methods, alternate measurements, and improved evidence-based guidelines. It is necessary for both researchers and practitioners to continue this pursuit.

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

Barton, E. E., Ferron, J. M., Kratochwill, T. R., Levin J. R., & Machalicek W. (August, 2014). Summer research training institute: Single-case intervention research design and analysis. Training session sponsored by the National Center for Special Education Research and the Institute of Education Sciences, Madison, WI.

Burns, M. K., Riley-Tillman, C., & Gibbons, A. M. (2013). *RTI applications volume 2: Assessment, analysis and decision making.* New York, NY: Guilford Press.

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., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review, 36*, 130–146.

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

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

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlational analysis for the behavioral sciences (2nd ed.) Hillsdale, NJ: Lawrence Erlabaum Associates.

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

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

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

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

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*, 184–192.

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

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

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

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

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

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

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.

Individuals With Disabilities Education Improvement Act, H. R. 1350, 108th Congress (2004).

Iowa Area Education Agencies. (2014). *Area education agency special education procedures.* Retrieved from http://www.iowaideainfo.org/vimages/shared/vnews/stories/4a8b1534597fd/Special%20Education%20Procedures%20July%201%202014.pdf

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

Kane, M. (2013). The argument-based approach to validation. School Psychology Review, 42, 448–457.

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.

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

Shinn, M. R., & Shinn, M. M. (2002). AIMSweb training workbook: Administration and scoring of Reading Curriculum Based Measurement (R-CBM) for use in general outcomes measurement. Available from www.aimsweb.com

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

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.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching.* Columbus, OH: C. E. Merrill.

Williams, B. B., & Hunley, S. (2015). *Using single-subject design to demonstrate positive impacts on student outcomes.* Workshop presented at the National Association of School Psychologists Annual Conference, Orlando, FL.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice, 18*, 137–146.

Date Received: December 15, 2014
Date Accepted: September 9, 2015
Associate Editor: Lisa Bowman-Perrott ■

Ethan R. Van Norman, PhD, is an assistant professor in the School Psychology program in the Department of Counseling and Psychological Services at Georgia State University. His research primarily focuses on evaluating and improving the technical adequacy of academic and behavioral measures used in schools. In addition, Dr. Van Norman conducts research aimed at building the capacity of educators and school psychologists to use data meaningfully to make sound educational decisions.

Theodore J. Christ, PhD, is a professor of school psychology in the Department of Educational Psychology and is the Director for the Center of Applied Research and Educational Improvement (CAREI) and Co-Director of the Research Institute for Problem Solving, which all reside at the University of Minnesota. It was his work in those roles that established him as the Founder and Chief Scientific Officer of FastBridge Learning (fastbridge.org). Dr. Christ is engaged to innovate techniques and technology to serve professional educators who seek to improve educational outcomes. He is interested and engaged with the content, methodology, analytics, technology, software, and people who pursue the same.