Evaluating Curriculum-Based Measurement from a Behavioral Assessment Perspective

Scott P. Ardoin, Claire M. Roof,
Cynthia Klubnick & Jessica Carfolite

Curriculum-based measurement Reading (CBM-R) is an assessment procedure used to evaluate students' relative performance compared to peers and to evaluate their growth in reading. Within the response to intervention (RtI) model, CBM-R data are plotted in time series fashion as a means modeling individual students' response to varying levels of instruction and the interpretation of these data is used as a source of information for making special education eligibility decisions. While substantial evidence exist demonstrating the reliability and validity of CBM-R procedures from a classical test theory (CTT) perspective, little evidence exist demonstrating the quality of CBM-R from a behavioral assessment perspective. This paper discusses (a) the necessity of evaluating CBM-R from a behavioral assessment perspective and (b) those studies which have evaluated CBM-R from a perspective other CTT, and (c) recommendations for future research.
Keywords: Curriculum-based measurement, behavioral assessment, absolute versus relative performance, accuracy, sensitivity, treatment utility.

Curriculum-based measurement for reading (CBM-R) procedures were developed in the late 1970's and early 1980's as a set of standardized assessment tools for gauging students' academic performance in reading. It was developed in order to provide teachers with an efficient, easily understood measurement system yielding relevant data about students' level of performance, as well as their reading growth over time (Deno, 1985). Educators and researchers have noted that an important characteristic of CBM-R is its ability to measure both inter-individual differences in groups of students as well as intra-individual change within specific students (Fuchs & Fuchs, 1998; Fuchs, Fuchs, & Speece, 2002). While CBM-R data were initially used exclusively to guide low-stakes educational decisions (Deno, 1985; Deno, 1986; Deno, Marston, & Tindal, 1985; Deno & Shinn, 1989), CBM-R data are now being used for making high-stakes decisions (i.e., special education eligibility) within Response to Intervention (RtI) models.

Several features distinguish CBM-R procedures from other standardized measures used to assess students' reading. First, the assessment materials are relatively cheap and it requires little time to administer probes to students. Second, CBM-R is meant to be a measurement of students' global reading performance, which allows for practitioners to evaluate how students' are progressing toward towards long-term goals (Deno, Fuchs, Marston, & Shin, 2001; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993). Finally, as described by Deno et al.

> "CBM-R departs from conventional psychometric applications by integrating the concepts of standardized measurement and traditional reliability and validity with features from behavioral and observational assessment methodology: repeated performance sampling, fixed time recording, graphic display of times-series data, and qualitative descriptions of performance" (Deno et al., 2001, p. 508)

These characteristics makes CBM-R ideal for use within an RtI model, as an instrument used to both identify students at-risk for academic problems and to evaluate individual students' response to instruction.

Within the Fuchs and Fuchs (1998) dual discrepancy RtI model, CBM-R data are first used to identify students who are at-risk for academic problems based upon a comparison of their performance to

that of their normative group (i.e., nomothetic context). Such decisions are relatively low-stakes decisions, given that the identification of a student at-risk simply results in the student being provided with supplemental instruction. CBM-R data modeling individual student's response to the supplemental instruction is then used as a primary source of data for making high-stakes decisions. Students' response to supplemental instruction is generally evaluated using CBM-R progress monitoring procedures, which entails the frequent administration of CBM-R probes and plotting of collected data in time-series fashion. Progress monitoring data are evaluated by comparing the plotted data to a pre-established goal line or an estimate of weekly growth calculated using ordinary least square regression techniques. Regardless of the method of comparison, these evaluations of data are within an idiographic context in which an individual's data are compared to his/her previous performance(s). Based upon the evaluation of data, one of the following high-stakes decisions is made: (a) the intervention was successful, therefore the student is not eligible for special education and the intervention should be terminated, (b) the intervention was not adequate and a more intense intervention is needed, or (c) supplemental interventions of varying levels of intensity have not been successful and the student is eligible for special education.

The importance of the psychometric adequacy of CBM-R is especially salient now that CBM-R data are being used as a primary source for making high-stakes educational decisions (Ardoin & Christ, in press; Christ & Ardoin, 2007). As previously noted, CBM-R combines features from CTT as well as behavioral assessment. Ample evidence is available demonstrating the reliability and validity of CBM-R from the CTT perspective. Fuchs, Tindal, and Deno (1984) reported that CBM-R probes demonstrated adequate criterion validity when compared to traditional measures of reading, such as the Stanford Diagnostic Reading Test and the Woodcock Reading Mastery Test. CBM-R has also demonstrated a high level of correspondence with other widely used reading tests (Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, & Maxwell, 1988; Shinn, 1989), tests of reading comprehension, and teachers' perceptions of reading competence (Fuchs & Fuchs, 1992). Furthermore, CBM-R data have been shown to effectively discriminate populations such as general and special education students (Deno, 1985; Deno, Shinn, Marston, & Tindal, 1983) and students from different grade levels (Deno, 1985; Fuchs et al., 1993).

Despite proponents of CBM-R frequently citing studies employing CTT as proof that CBM-R is psychometrically sound and thus appropriate for use within an RtI framework, CTT is not necessarily the most appropriate model for evaluating its adequacy within an RtI framework. Fuchs and Fuchs (1998) note that traditional psychometric standards provide insufficient evidence for using a measure to model individual growth. CTT evaluates the stability of group level data in which it is ideal for each student's performance to remain stable relative to the sampled group. Any changes in performance relative to the sampled group are treated as error and either explained away or ignored. Observed scores are considered to represent an individual's true score plus error. In contrast to these principals of CTT, when evaluating CBM-R data plotted in time series fashion (a) data are being compared to the previous performances of the same student, rather than to a sample group; and (b) change in student performance is desired, rather than treated as error. It is the departure of CBM-R from traditional psychometrics into behavioral assessment that permits modeling of an individual student's growth and thus it is also necessary that CBM-R be evaluated from a behavioral assessment perspective in order to demonstrate its appropriateness within an RtI framework.

Evaluation of CBM-R from a Behavioral Assessment Perspective

Nelson and Hayes (1979) state that the goal of behavioral assessment is to understand and modify behavior through the identification of meaningful behavioral responses and the environmental and internal factors that control them. Practitioners dealing with behavioral issues are motivated by the utility of measurement in generating change in all relevant environments, and are therefore interested in what makes an individual's behavior fluctuate from situation to situation. For these reasons, behavioral assessment is conducted from an idiographic approach in which the individual is perceived as an entity

with qualities that distinguish him or her from others, rather than as part of a cohort or category. Due to the fundamental differences in the nature of the traditional and behavioral approaches to assessment, Hayes, Nelson, and Jarrett (1986) maintain that the two areas should have separate standards for evaluation. Hayes et al. specified three distinct areas in which behavioral assessment tools should be evaluated: accuracy, sensitivity, and treatment utility.

Accuracy. Cone (Cone, 1981) defined accuracy as "how faithfully a measure represents objective topographic features of a behavior of interest" (p. 59). The first step towards establishing the accuracy of a measure is to (a) develop explicit instructions to accompany the measure in order to establish consistency in measurement and (b) select a behavior that is verbally definable (Cone, 1977, 1992; Nelson & Hayes, 1979). Accuracy of a measure can then be established by comparing measurement results to those of some standard (similar to criterion-related validity). Since the targets of behavioral assessment are directly observed, accuracy is generally construed as how a measure compares with "reality" (Hayes et al., 1986). This definition contrasts with the interchangeability of the concepts of accuracy and reliability (i.e., consistency among scores across time, settings, items, etc.) that is often encountered in the literature on traditional assessment, where the focus is on inferred traits (Hayes et al., 1986). Accuracy of a behavioral measure can be established by weighing responses against those obtained with a mechanical record, or by emitting a scripted set of responses to ensure that the measure accurately detects them (Hayes et al., 1986).

Accuracy of CBM-R. CBM-R procedures clearly address the first steps towards establishing accuracy (i.e., explicit instructions, verbally definable target behavior). Explicit administration and scoring procedures have been available since the early development of CBM-R. These procedures involve having students read a passage for one minute while an examiner records misread words and words which the examiner provides to students when they hesitate on a word for three seconds. A word is marked as misread if a student mispronounces a word given the context of the story, skips a word, or transposes two words. Examiners provide words on which students hesitate for three seconds, but examiners do not correct students if they misread a word. After one minute elapses, student are asked to stop reading and the examiner calculates words read correctly by subtracting misread words from total words read. Evidence of the importance of adhering to the explicit administration directions has been provided by studies which have found variation in student performance as a function of directions given to students (Eckert, Dunn, & Ardoin, 2006; Fuchs, Tindal et al., 1984). Fortunately, researchers have consistently used the same explicit administration and scoring procedures across studies evaluating the reliability and validity of CBM-R.

Despite having clearly addressed the first steps of establishing the accuracy of a measure, researchers have not established the accuracy of CBM-R by comparing it to a known standard. Stoner (1992) suggested that accuracy was a "questionable metric" for evaluating CBM-R because no "gold standard" has been developed against which these data can be compared. Cone (1992) proposed using a tape of words read into a recorder at a speed paced by a metronome or using a computer with an audio digitizer that could count the number of words read at different speeds as a means of independently evaluating the accuracy with which examiners record words read correctly per minute (WRCM). Fifteen years later, researchers have not adhered to the suggestions made by Cone. Instead, the accuracy of CBM-R observations is based upon assessments of inter-rater agreement. Using audio recorders, numerous studies have evaluated the accuracy of observers on a word by word basis, comparing each word the observer reported as correct and incorrect to each word a blind rater scored as correct and incorrect. Inter-rater agreement is calculated as the percentage of agreement over disagreements and generally exceeds 90% (e.g., Ardoin, Suldo, Witt, Aldrich, & McDonald, 2005; Ardoin et al., 2004; Graney & Shinn, 2005). High rates of inter-rater reliability resulting from an easily quantified target behavior and standardized instructions, administration, and scoring procedures, support the measure's capacity to accurately detect and report the behavior of interest.

Instead of evaluating accuracy using procedures recommended by Cone, accuracy within the CBM-R literature typically is referred to in terms of standard error of measurement (*SEM*). *SEM* is the amount of error associated with the measurement method, and is influenced by the quality of instrumentation (e.g., equivalence in level of difficulty for parallel probes), as well as administration conditions (Christ & Ardoin, 2007). A method for evaluating this form of accuracy is to conduct similar but independent measurements (Hayes et al., 1986) in a short period of time during which meaningful growth is unlikely to have occurred (Ardoin & Christ, in press; Christ & Ardoin, 2007; Poncy, Skinner, & Axtell, 2005). An examiner may accurately record the WRCM for a student in several administrations of parallel probes during this time; yet we would still expect scores to be somewhat variable if, for example, the probes used are inconsistent in difficulty. Although consumers of CBM-R data rarely report the standard error associated with scores, it has been found that the *SEM* associated with CBM-R performance level estimates in typical testing conditions is actually quite substantial, approximating 10 WRCM (Christ & Silberglitt, 2007). This number is significant when considering that the expected performance level of a student in the spring of the first grade is only around 40 WRCM (Good, Simmons, & Kame'enui, 2001). Due to this performance variability, it has been recommended that the median score from a minimum of three probes administered at once should be used for screening purposes (Ardoin & Christ, in press; Poncy et al., 2005).

Because the majority of research evaluating CBM-R has been conducted from a CTT perspective, in which error is typically either explained away or ignored, only a few studies have evaluated the sources of error associated with student performance. In one of the few studies specifically evaluating CBM-R from a behavioral perspective, Derr-Minneci and Shapiro (1992) investigated the potential sources of error related to various environmental factors and found main effects for setting, tester, and timed versus untimed administrations. More recently efforts have been made to quantify the sources of error associated with CBM-R. Using generalizability theory researchers have separated the variance in CBM-R scores into that which is attributed to the person (i.e., student learning), the item (i.e., CBM-R probe passages), and residual error (Christ & Ardoin, 2007; Poncy et al., 2005). Results of Poncy, et al. indicated that up to 19% of the total variance in students' performance level could be explained by either the item (up to 10%) or residual error. Studies by Ardoin and colleagues (Ardoin & Christ, in press; Ardoin et al., 2005; Christ & Ardoin, 2007) suggest that the inability of readability formulas to adequately control passage difficulty is likely the primary source of variance (i.e., error) associated with items.

Depending on the context in which data are being evaluated, variability in student performance as a function of variation in passage difficulty can lead to misinterpretations of observed performance. For instance when evaluating relative performance, a student's WRCM is compared to other students administered the same probe(s) and thus the difficulty of the probe will be constant across students resulting in little to no change in relative performance. However, if performance is being evaluated at the individual child level, a student's performance on a probe is likely to be compared to a pre-established criterion or to the child's previous/future performances. Neither comparison considers the possibility of variation in passage difficulty, which is likely to result in variation in student performance that is not a function of change in the student's global reading skills. Given variability in student performance as a function of tester, setting, and passage difficulty, it is important that schools recognize that each observed score accounts for a student's true score plus error. In recognizing the lack of precision of an individual's performance confidence intervals should be placed around observed student CBM-R performance (Christ & Coolong-Chaffin, 2007).

Sensitivity. Psychometric theory cannot be unequivocally applied to behavioral assessment because behavior is more specific to situations than it is consistent across situations (Mischel, 1968). Inconsistencies in results could be due to actual changes in behavior, not weaknesses of the measure (Nelson & Hayes, 1979). This issue marks a key difference between the behavioral approach to assessment and its traditional counterpart (i.e., CTT) in which consistency is highly prized and

performance variability is labeled "error" (Cone, 1977; Hayes et al., 1986). A measure designed to detect a particular behavior should be sensitive to factors expected to have an impact on the occurrence of that behavior; without such consideration, concerns about the accuracy, reliability, or validity of the behavioral assessment tool are rendered "meaningless" (Hayes et al., 1986, p. 493).

In effect, the analysis of environmental influences on behavior constitutes the premise of behavioral assessment (Derr-Minneci & Shapiro, 1992; Nelson & Hayes, 1979). Nevertheless, sensitivity can be viewed as a double-edged sword precisely because the assessment process plays an integral role in the portion of variance that is due to the "situation" (Nelson & Hayes, 1979). For example, while an educator would want a test of academic performance to be sensitive to changes in learning, it would generally not be desirable if performance was affected by variables such as characteristics of the administrator. It is therefore imperative to identify and eliminate sources of unnecessary and uncontrollable systematic or random error and to foster an awareness of the potential causes of variation in assessment results (Christ, 2006; Christ & Ardoin, 2007; Hayes et al., 1986).

There are several means by which the sensitivity of a behavioral measure can be evaluated (Hayes et al., 1986). One method is to conduct idiographic correlations between different measures of the same construct over time. This technique allows for the recognition of changes in data at the *individual* level, taking into account the error associated with testing over time for each participant. Another method to evaluate the sensitivity of a measure is to implement an intervention that theoretically should result in changes in behavior and then to evaluate whether the measure detects differences in behavior (e.g., Ardoin & Martens, 2004). Additionally, a criterion measure known to demonstrate adequate sensitivity (i.e., a gold standard) can be used for comparison purposes (Hayes et al., 1986). When assessing sensitivity using these procedures, small-n-design procedures must be employed because the issue of sensitivity is often temporal in nature (i.e., involves an idiographic comparison of an individual's performance to his or her past performances). Understanding the nature of the data at the individual level is critical when the data are to be used for making decisions within an idiographic context (i.e., comparing a person's to his/her past performances). Nomothetic or group-level analyses, in contrast, do not take into account the fact that situational factors affect individuals differentially.

Sensitivity of CBM-R. Several intervention based studies provide evidence of the sensitivity of CBM-R by demonstrating that effective interventions can result in changes in a student's level of performance on a specific probe. Using multi-element designs, researchers have found that students performance on intervention passages and generalization passages increases as a function of varying degrees of intervention intensity when evaluated using CBM-R procedures (Ardoin, McCall, & Klubnik, 2007; Eckert, Ardoin, Daisey, & Scarola, 2000; McCurdy, Daly, Gortmaker, Bonfiglio, & Persampieri, 2007). While these studies demonstrate that CBM-R procedures are sensitive to direct intervention effects, they fail to provide evidence that CBM-R data used to model student growth are sensitive to variations in instruction.

Evidence used to support CBM-R as a measure of global gains in student achievement across time has largely centered on the idea that a steeper slope for WRCM signifies a more sensitive measure. From this perspective, when compared to traditional achievement measures (e.g., norm referenced individually/group administered achievement tests), CBM-R progress-monitoring techniques are clearly more sensitive to fluctuations in student learning (Deno, 1985; Fuchs, Fuchs, & Hamlett, 1989a; Marston, Fuchs, & Deno, 1986). While traditional reading achievement tests have been unable to differentiate between low-achieving and students with learning disabilities, CBM-R does in fact show a meaningful difference in the academic gains of these two groups of students (Shinn, Ysseldyke, Deno, & Tindal, 1986).

Studies have also been conducted demonstrating the sensitivity of CBM-R to model the growth of students by demonstrating differences in rates of growth as a function of variables that theoretically should result in different rates of gain. Some of the first studies in the area of the sensitivity of CBM-R involved longitudinal research on the relative effectiveness of general education and special education services for students at risk for reading problems (Marston, 1987). Weekly progress-monitoring data were used to show that academic growth, on average, was higher in the special education settings. Researchers have also used CBM-R to compare the effects of instruction in different classrooms, providing evidence that the quality of instruction provided to a child is a critical consideration in the evaluation of student performance (Fuchs & Fuchs, 1998; Speece & Case, 2001). Studies have also been conducted evaluating the sensitivity of CBM-R to different methods of reading instruction (literature-based vs. traditional based). Hintze, Shapiro, and Lutz (1994) found that on average CBM-R slopes were greater for the group of students instructed in the traditional curricula. It was concluded that steeper slopes may have signified the greater sensitivity of CBM-R to learning in traditional curricula, possibly as a result of greater overlap between the traditional curricula and assessment materials. Further evidence of the sensitivity of CBM-R progress monitoring data to variables that theoretically should alter rates of growth is provided by Allinder and Eicher (1994). In this study a decline in rates of growth was detected during the summer months due to lack of active instruction.

While results of the above cited studies (e.g., Allinder & Eicher, 1994; Hintze et al., 1994; Shinn et al., 1986) provides some evidence of the sensitivity of CBM-R progress monitoring procedures used to model students global gains in reading, it is important to attend to the context in which these studies have been conducted. These studies have evaluated the sensitivity of CBM-R to global gains in reading using CTT procedures to examine the growth of groups of students. Analyses of students' CBM-R progress monitoring data in order to evaluate their response to instruction within an RtI framework are not made based upon group level data. Rather, analyses within an RtI framework are conducted within an idiographic context to determine whether across time the observed performances of an individual student increases, decreases, or stays relatively the same.

Results of recent studies bring into question whether CBM-R might in fact be too sensitive to environmental factors for high-stakes decisions to be made with confidence outcomes (Christ, 2006; Christ & Ardoin, 2007; Christ & Silberglitt, 2007; Poncy et al., 2005). Variability in student performance as a function of inconsistencies in probe difficulty negatively impacts the reliability, accuracy, stability, and thus sensitivity of CBM-R progress monitoring (Christ, 2003; Hintze, Owen, Shapiro, & Daly, 2000; Hintze & Shapiro, 1997; Hintze, Shapiro, & Daly III, 1998; Hintze et al., 1994). Using estimates of error associated with rates of growth calculated for individual students from previously published studies, Christ (2006) found that the magnitude of error often exceeds expected rates of growth. These results indicate that CBM-R procedures are extremely sensitive to variables other than changes in rates of growth, suggesting that high-stakes decisions based upon estimates of CBM-R progress monitoring data should be made with extreme caution.

The behavioral assessment literature does offer several directions for evaluating data used in an idiographic context that should be applied to CBM-R. For example researchers could conduct idiographic correlations (Hayes et al., 1986) using two sets of CBM-R probe sets. Although significant correlations between individual students' performances on the two probe sets would not necessarily indicate that daily change was a function of growth, the data would provide evidence that changes were systematic and not a function of random error. Additional research must be conducted to identify and eliminate sources of error associated with CBM-R progress monitoring data to ensure that change in student performance is a function of change in global reading achievement.

Treatment Utility. Treatment utility was selected as an area of appraisal for behavioral assessment because the objective of behavioral assessment is to inform intervention (Hayes et al., 1986). Quality

assessment measures help practitioners develop goals that allow them to predict and control behavior in the most efficient manner possible, ultimately enhancing treatment outcomes (Nelson & Hayes, 1979). The first step towards establishing the treatment utility of a measure is selecting a target behavior that has meaning and can efficiently discriminate between individuals who need intervention and those who do not. Once a target behavior and a procedure for detecting it have been selected, the best way to evaluate treatment utility is to consider the results of decisions based on assessment data (Messick, 1980). This can be done after assessing the effectiveness of a treatment strategy by using a single-case experimental design with many subjects and correlating individual assessment results with individual treatment effects. Hayes et al (1986) and Nelson and Hayes (1979) provide several examples of idiographic methods for evaluating treatment utility when the relationship between assessment and treatment outcomes are predicted a priori. For example the "manipulated assessment" is described as systematically varying an aspect of assessment (e.g., the target behavior or the assessment method), and implemented treatment according to the assessment data. Treatment outcomes are then evaluated to assess the effectiveness of each. In a second method ("manipulated use"), all participants receive the same assessment, but the correspondence between assessment and treatment is manipulated. Finally, the "observed differences" method involves all participants receiving the same treatment which is evaluated for effectiveness, but differences in assessment are noted (Hayes et al., 1986; Nelson & Hayes, 1979).

Treatment Utility of CBM-R. CBM-R, consistent with a behavioral approach to assessment, was developed precisely for the purpose of providing information that teachers could use to better understand how a child functions in the curriculum (Deno, 1985). CBM-R procedures inform teachers by discriminating between typical and atypical performance (Deno, 2002; Shinn, Thomas, & Grimes, 2002), thus illuminating a student's need for intervention. Once a student is determined to be in need of academic assistance, CBM-R is designed to show the student's progress while receiving intervention (Fuchs, 2003), this information is helpful in making instructional decisions. Numerous studies have been conducted to investigate the ability of CBM-R to enhance teacher planning and most importantly improve student achievement (Fuchs, Deno, & Mirkin, 1984; Fuchs, Fuchs, & Hamlett, 1989b; Fuchs, Fuchs, Hamlett, & Stecker, 1991).

Evidence of the treatment utility of CBM-R is often provided by studies in which the students of teachers who view their students CBM-R progress monitoring data make greater academic gains than those students whose teachers do not view CBM-R progress monitor data. Fuchs and Fuchs (1986) conducted a meta-analysis of this research and found an effect size of .7 for progress-monitoring. In one study included in this review, Fuchs, Deno, & Mirkin (1984) investigated the effect of using CBM-R during the course of 18 weeks. A single CBM-R probe was administered twice weekly to each student in the experimental group. Teachers graphed data and used graphed data as a means of determining whether students were making progress toward IEP goals and thus whether instructional modifications were needed. In contrast, control group teachers used traditional informal procedures to track student growth (i.e., a "manipulated assessment" design). Results revealed increased student goal revisions initiated by the teacher, instructional structure, student awareness of their own learning, and student reading achievement for those students and teachers in the experimental condition. This study provided evidence supporting teachers' use of CBM-R when accompanied with explicit guidelines for modifying instruction. Moreover, Wesson (1991) found that special education students, whose teachers referred to CBM-R progress monitoring data and participated in group consultation regarding student performance, achieved higher levels of reading achievement than those students for whom progress-monitoring data were not collected.

Similar studies, conducted in general education settings where teachers are typically allotted fewer resources for attending to individual students, have found conflicting results. Graney and Shinn (2005) examined whether CBM-R progress monitoring information provided to a teacher about individuals and groups of students had an effect on reading performance. At five weeks of progress-

monitoring, teachers in the experimental groups viewed trend lines for their students and were provided the opportunity to discuss possible ways to adapt instruction with a consultant. Students in the control condition were tested with CBM-R, but teachers did not view these data (i.e., "manipulated use" design). Results indicated that feedback (concerning individuals or entire classrooms) had no impact on the reading achievement of the students at the group level, and that individual feedback may have actually had a negative effect. It appears that providing information to general education teachers on students' slopes, and even altering instruction based on these slopes, may not be enough to effect positive student change. This evidence is supported by other findings showing that when teachers are provided with CBM-R data without specific guidelines on how to use this information to guide instruction (i.e., "change the program decision rules"), the effects on student learning are negligible (Fuchs et al., 1989a).

Consistent with those studies evaluating the accuracy and sensitivity of CBM-R, studies evaluating the treatment utility of CBM-R at the individual student level are sparse. Although studies by Daly and colleagues (Daly, Persampieri, McCurdy, & Gortmaker, 2005; McCurdy et al., 2007) provide evidence of the use of CBM-R for measuring the effects of brief instructional trials on individual students, researchers have not yet evaluated the impact of progress monitoring on individual students. Caution should therefore be taken when using CBM-R progress monitoring data to inform treatment and decisions about individual students' responsiveness to instruction within an RtI framework. Within-subject design studies must be conducted correlating individual assessment results with individual treatment outcomes in order to account for error at this level of analysis. An additional issue that should be considered in future research examining the treatment utility of CBM-R progress monitoring data is whether improvements in performance over control students are in fact a function of making decisions based upon progress monitoring data. While most previous studies have employed a control condition for comparison purposes, these studies did not use a yoked control group whose instruction was changed on the basis of a student in the progress monitoring condition. It is possible that merely being exposed to dynamic instruction characterized by frequent changes benefitted students academically. Adding a yoked condition with these characteristics would greatly enhance the strength of the "manipulated assessment" and "manipulated use" designs (Hayes et al., 1986).

Conclusions

Numerous studies have been conducted for the purpose of exhibiting the strengths of CBM-R as an innovative academic assessment tool, and the results of these studies are often used to applaud CBM-R for the improvements that it represents over the traditional IQ/achievement test discrepancy model. Influenced by the proliferation of the RtI for determining special education eligibility, the educational context in which CBM-R is used has evolved to the present state, where both low-stakes and high-stakes decisions are made about individual students based on time-series data. Abundant evidence has been provided for the psychometric adequacy of CBM-R from a CTT perspective supporting its many uses for making low-stake educational decisions for students. Substantially fewer studies have however been conducted demonstrating the quality of CBM-R from a behavioral assessment perspectives in which the quality of a measure is based upon its accuracy, sensitivity, and treatment utility for making decisions within an idiographic context. It is within this context that high stake decisions are being made regarding student's special educational eligibility.

In order for CBM-R data to be used as a primary source of information for high-stakes decision making several steps must be taken. First, equivalent sets of CBM-R reading probes must be developed. We know that readability formulas are inadequate for selecting passages equivalent in level of difficulty and that variation in student performance is a primary source of error. Ignoring this fact will result in students being misidentified for special education. Second, we must begin to identify sources of error other than variability in passage difficulty and determine procedures for minimizing this error. Considering studies have illustrated that variation in CBM-R directions, performance feedback, goal

setting, and reinforcement influences student performances these may be areas to consider (Eckert et al., 2006; Fuchs, Tindal et al., 1984). For example, standard CBM-R directions simply inform students to do their "best reading." It is possible that the meaning of "best reading" changes as students observe themselves being timed repeatedly (Colon & Kranzler, 2006). Defining best reading for students and providing them with a constant source of reinforcement of doing their best reading might results in more consistent performance. Finally, researchers need to address the question of how many data points across, how many weeks are needed to make a reliable and accurate decision regarding the effectiveness of instruction for students. Current sources used to provide evidence that 20 data points are sufficient for making decisions based upon CBM-R data were designed for making low-stake decisions (Good & Shinn, 1990; Shinn, Good, & Stein, 1989). It is also questionable whether these studies actually support the use of 10-20 data points as the magnitude of error in predicting future student performance within these studies is greater than expected rates of student gain (Good & Shinn, ; Shinn et al.), which is consistent with recent research (Christ, 2006).

The purpose of this manuscript was to evaluate CBM-R from a behavioral perspective and thus to highlight the fact that there is a considerable amount of work that needs to be conducted before decisions using CBM-R data can be made with great confidence. The intention of this paper was not, however, to suggest that CBM-R should not be used within an RtI framework. CBM-R data are invaluable for schools to evaluate the quality of instruction for their student body and to identify students in need of supplemental instruction. CBM-R data are also useful as one component of multiple sources of assessment data that can be used for evaluating a student's response to instruction. Schools should, however, consider alternative means of assessing intervention effects that directly evaluate the impact of instruction; this might allow decisions regarding intervention effects to be made within relatively short periods of time (e.g., 5 weeks). It is essential to remember that CBM-R is a global measure of reading achievement. Another way to explain this is that CBM-R evaluates generalization effects. In the same way that the effects of an intervention targeting classroom behavior should not be evaluated by only evaluating generalization effects, we should not only evaluate the effects of a reading intervention by evaluating generalization effects. Generalization does not occur naturally and it is likely that generalization effects will require longer to appear than schools allot for evaluating intervention effects. Measures must be used that evaluate whether a student is mastering the individual skills being taught. Only with mastery of component skills is generalization likely to occur, resulting in improvements in global/composite skills (Ardoin & Daly, 2007; Binder, 1996).

References

Allinder, R. M., & Eicher, D. D. (1994). Bouncing back: Regression and recoupment among students with mild disabilities following summer break. *Special Services in the Schools, 8*(2), 129-142.

Ardoin, S. P., & Christ, T. J. (in press). Evaluating curriculum-based measurement slope estimates using data from tri-annual universal screenings. *School Psychology Review*.

Ardoin, S. P., & Daly, E. J., III. (2007). Introduction to the Special Series: Close Encounters of the Instructional Kind--How the Instructional Hierarchy is Shaping Instructional Research 30 Years Later. *Journal of Behavioral Education, 16*(1), 1-6.

Ardoin, S. P., & Martens, B. K. (2004). Training children to make accurate evaluations: Effect on behavior and the quality of self-ratings. *Journal of Behavioral Education, 13*, 1-23.

Ardoin, S. P., McCall, M., & Klubnik, C. (2007). Promoting Generalization of Oral Reading Fluency: Providing Drill versus Practice Opportunities. *Journal of Behavioral Education, 16*(1), 55-70.

Ardoin, S. P., Suldo, S. M., Witt, J. C., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly, 20*(1), 1-22.

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connel, J. E., Koenig, J. L., Resetar, J. L., et al. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probe when conducting universal screening. *School Psychology Review, 33*, 218-233.

Binder, C. (1996). Behavioral fluency: Evolution of a new paradigm. *The Behavior Analyst, 19*, 163-197.

Christ, T., J. (2003). The effects of passage-difficulty on CBM progress monitoring outcomes: Stability and accuracy. *Dissertation Abstracts International: Section B: The Sciences & Engineering, 62*(7), 3414.

Christ, T., J. (2006). Short term estimates of growth using curriculum-based measurement of oral reading fluency: Estimates of standard error of the slope to construct confidence intervals. *School Psychology Review, 35*(1), 128-133.

Christ, T. J., & Ardoin, S. P. (2007). Curriculum-based measurement: Passages difficulty. *Manuscript submitted*.

Christ, T. J., & Coolong-Chaffin, M. (2007). Interpretations of curriculum-based measurement outcomes: Standard error and confidence intervals. *School Psychology Forum: Research in Practice, 1*(2), 75-86.

Christ, T. J., & Silberglitt, B. (2007). Curriculum-based measurement of oral reading fluency: The standard error of measurement. *School Psychology Review*, (36), 130-146.

Colon, E. P., & Kranzler, J. H. (2006). Effect of Instructions on Curriculum-Based Measurement of Reading. *Journal of Psychoeducational Assessment, 24*(4), 318-328.

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavior Therapy, 8*, 411-426.

Cone, J. D. (1981). Psychometric considerations. In M.Herson & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (2nd ed., pp. 38-68). Elmsford,NY: Pergamon Press.

Cone, J. D. (1992). Accuracy and curriculum-based measurement. *School Psychology Quarterly, 7*(1), 22-26.

Daly, E. J., III, Persampieri, M., McCurdy, M., & Gortmaker, V. (2005). Generating reading interventions through experimental analysis of academic skills: Demonstration and empirical evaluation. *School Psychology Review, 34*(3), 395-414.

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*(3), 358-374.

Deno, S. L. (2002). Problem-solving as "best practices." In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (Vol. 1, pp. 37-56). Washington, DC: National Association of School Psychologists.

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

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

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36-45.

Deno, S. L., Shinn, J., Marston, D., & Tindal, G. (1983). Oral reading fluency: A simple datum for scaling reading disability. *Topics in Reading and Reading Disabilities, 2*, 53-59.

Deno, S. L., & Shinn, M. R. (1989). *Curriculum-based measurement and special education services: A fundamental and direct relationship*. New York, NY, US: Guilford Press.

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

Eckert, T. L., Ardoin, S. P., Daisey, D. M., & Scarola, M. D. (2000). Empirically evaluating the effectiveness of reading interventions: The use of brief experimental analysis and single case designs. *Psychology in the Schools, 37*, 463-473.

Eckert, T. L., Dunn, E. K., & Ardoin, S. P. (2006). The effects of alternate forms of performance feedback on elementary-aged students' oral reading fluency. *Journal of Behavioral Education, 15*(3), 148-161.

Fuchs, D., & Fuchs, L. S. (1992). Limitations of a feel-good approach to consultation. *Journal of Educational and Psychological Consultation, 3*(2), 93-97.

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

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on student achievement, pedagogy, and student awareness of learning. *American Educational Research Journal, 21*, 449-460.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children, 53*(3), 199-208.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research and Practice, 13*(4), 204-219.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989a). Effects of alternative goal structures within curriculum-based measurement. *Exceptional Children, 55*(5), 429-438.

Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989b). Monitoring reading growth using student recalls: Effects of two teacher feedback systems. *Journal of Educational Research, 83*(2), 103-110.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teaching planning and student achievement in mathematics operations. *American Educational Research Journal, 28*(3), 617-641.

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

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28.

Fuchs, L. S., Fuchs, D., & Speece, D. L. (2002). Treatment validity as a unifying construct for identifying learning disabilities. *Learning Disability Quarterly, 25*(1), 33-45.

Fuchs, L. S., Tindal, G., & Deno, S. L. (1984). Methodological issues in curriculum based reading assessment. *Diagnostique, 9*, 191-207.

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

Good, R. H., III, Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision making utility of continuum of fluency-based indicators of foundational reading skills for third grade high-stakes outcomes. *Scientific Studies of Reading, 5*(2), 257-288.

Graney, S. B., & Shinn, M. R. (2005). Effects of Reading Curriculum-Based Measurement (R-CBM) Teacher Feedback in General Education Classrooms. *School Psychology Review, 34*(2), 184-201.

Hayes, S. C., Nelson, R. O., & Jarret, R. B. (1986). Evaluating the quality of behavioral assessment. In R. C. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 463-503). New York: Guilford.

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J., III. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15*(1), 52-68.

Hintze, J. M., & Shapiro, E. S. (1997). Curriculum-based measurement and literature-based readings: Is curriculum-based measurement meeting the needs of changing reading curricula? *Journal of School Psychology, 35*, 351-375.

Hintze, J. M., Shapiro, E. S., & Daly III, E. J. (1998). An investigation of the effects of passage difficulty level on outcomes of oral reading fluency progress monitoring. *School Psychology Review, 27*(3), 433-445.

Hintze, J. M., Shapiro, E. S., & Lutz, J. G. (1994). The effects of curriculum on the sensitivity of curriculum-based measurement in reading. *The Journal of Special Education, 28*, 188-202.

Marston, D. (1987). The effectiveness of special education: A time series analysis of reading performance in regular and special education settings. *The Journal of Special Education, 21*(4), 13-26.

Marston, D., Fuchs, L. S., & Deno, S. I. (1986). Measuring pupil progress: A comparison of standardized achievement tests and curriculum-related measures. *Diagnostique, 11*, 71-90.

McCurdy, M., Daly, E., Gortmaker, V., Bonfiglio, C., & Persampieri, M. (2007). Use of Brief Instructional Trials to Identify Small Group Reading Strategies: A Two Experiment Study. *Journal of Behavioral Education, 16*(1), 7-26.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012-1027.

Mischel, W. (1968). *Personality and Assessment.* Hoboken, NJ: John Wiley & Sons Inc.

Nelson, R. O., & Hayes, S. C. (1979). Some current dimensions of behavioral assessment. *Behavioral Assessment, 1*, 1-16.

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*(4), 326-338.

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

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

Shinn, M. R., Thomas, A., & Grimes, J. (2002). *Best Practices in Using Curriculum-Based Measurement in a Problem-Solving Model.* Washington, DC, US: National Association of School Psychologists.

Shinn, M. R., Ysseldyke, J. E., Deno, S. L., & Tindal, G. A. (1986). A comparison of differences between students labeled learning disabled and low achieving on measures of classroom performance. *Journal of Learning Disabilities, 19*(9), 545-552.

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology, 93*(4), 735-749.

Stoner, G. (1992). Validating curriculum-based measurement: Essential concerns from a behavioral perspective. *School Psychology Quarterly, 7*(1), 17-21.

Wesson, C. L. (1991). Curriculum-based measurement and two models of follow-up consultation. *Exceptional Children, 57*(3), 246-256.

Author Contact Information

Scott P. Ardoin, Ph.D.
Assistant Professor, School Psychology
University of South Carolina
Department of Psychology
Columbia, SC 29208
spardoin@sc.edu

803 777-7616

Claire M. Roof
USC School Psychology Graduate Student
University of South Carolina
Department of Psychology
Columbia, SC 29208
RoofCM@gwm.sc.edu
803 730-1210

Cynthia Klubnik
M.S. Psychology from Univ. of South Carolina
University of South Carolina
Department of Psychology
Columbia, SC 29208
klubnik@mailbox.sc.edu
860 798-9560

Jessica C. Williams
USC School Psychology Graduate Student
University of South Carolina
Department of Psychology
Columbia, SC 29208
jcarfolite@yahoo.com
803 422-9837
