

## A Longitudinal Examination of the Diagnostic Accuracy and Predictive Validity of R-CBM and High-Stakes Testing

John M. Hintze  
*University of Massachusetts*

Benjamin Silberglitt  
*St. Croix River Education District*

*Abstract.* The purpose of this study was to compare different statistical and methodological approaches to standard setting and determining cut scores using R-CBM and performance on high-stakes tests. One thousand seven hundred and sixty-six students were followed longitudinally from first through third grades using R-CBM benchmark assessment. In addition, students were administered the Minnesota Comprehensive Assessment (MCA) at the end of third grade. Predictive validity and diagnostic accuracy analyses using discriminative analysis, logistic regression, and receiver operator characteristic (ROC) curves were conducted. Results suggested that R-CBM is strongly associated with MCA performance at each grade level and is both accurate and efficient in predicting those students who are likely to pass the reading portion of the MCA beginning in first grade. Applied and theoretical implications are discussed along with future research.

Curriculum-based measurement (CBM) is a standardized set of measurement techniques used to index student academic performance in the basic skill areas of reading, mathematics, spelling, and written expression (Deno, 1985; Deno, Mirkin, & Chiang, 1982; Fuchs & Deno, 1991; Shinn, 1989b). As a variant of curriculum-based assessment (CBA), CBM uses the general education curriculum as the basis for test development and is designed primarily as a measurement and evaluation system that school psychologists and teachers can

routinely use to monitor individual student progress and instructional effectiveness.

CBM differs from other forms of CBA in a number of important ways (Fuchs & Deno, 1991). First, the focus of CBM is on broad long-term goal objectives, rather than short-term objectives. With CBM, practitioners specify what they want students to achieve by year's end or longer. These long-term objectives structure the assessment process throughout the progress monitoring period, as the same performance objective is continually assessed.

---

Sincere appreciation is extended to the AIMSweb Science Committee for their thoughts and suggestions regarding this study.

Correspondence concerning this article should be addressed to John M. Hintze, PhD, University of Massachusetts at Amherst, School Psychology Program, 362 Hills South, Amherst, MA 01003; E-mail: [hintze@educ.umass.edu](mailto:hintze@educ.umass.edu)

Copyright 2005 by the National Association of School Psychologists, ISSN 0279-6015

Focusing on the broad goals of the curriculum rather than a series of short-term objectives allows CBM to attend to the assessment of more general integrated outcomes as they occur in context. The result of such measurement focuses the attention of the assessment on the broader desired outcome of instruction. This is in contrast to mastery or criterion-referenced approaches whereby the assessment material changes with each new short-term objective requiring the curriculum to be decomposed and compartmentalized for assessment. Second, because it focuses on broad aspects of the curriculum, CBM allows for the assessment of retention and generalization of learning. Using a domain sampling approach to test development, CBM draws on a broad domain of skills representing the current instructional focus, as well as those representing past and future instructional targets. As a result, CBM produces performance indicators that assess current learning in addition to the retention and generalization of previously mastered material. A third distinguishing feature of CBM is that it specifies the measurement and evaluation procedures to be used, including methods for generating test stimuli, administering and scoring tests, and summarizing and making inferences from the data collected. This again is in contrast to other forms of CBA where the administration and scoring, as well as test development procedures, are not standardized and are left up to the will of the examiner. Using standardized administration and scoring procedures allows for comparison of scores across students, as well as the comparison of scores within-student across time.

A substantial research literature has developed to demonstrate that CBM can be used effectively to gather student performance data to support a wide range of educational decisions (Deno, 2003). For example, CBM has been shown effective in improving instructional programs and enhancing teacher instructional planning through the use of goal setting, progress monitoring, and evaluating the effects of changes in a formative evaluation model (Fuchs, Deno, & Mirkin, 1984; Fuchs, Fuchs, & Hamlett, 1993; Fuchs, Fuchs, Hamlett, & Stecker, 1991); developing local normative

performance standards (Marston & Magnusson, 1988; Shinn, 2002); screening to identify students academically at risk (Deno, Reschly-Anderson, Lembke, Zorka, & Callender, 2002; Marston & Magnusson, 1988); evaluating classroom prereferral interventions (Shinn, 1995; Tilly & Grimes, 1998); offering alternative special education identification procedures (Fuchs & Fuchs, 1998; Marston & Magnusson, 1988; Marston, Mirkin, & Deno, 1984; Shinn, 1989a); and recommending and evaluating inclusion (Fuchs, Roberts, Fuchs, & Bowers, 1996; Powell-Smith & Stewart, 1998).

With increasing attention to accountability and high-stakes assessment, the relationship between CBM and state-mandated testing programs has been a recent topic of interest. In particular, the criterion and predictive validity of reading CBM (R-CBM) has been examined as the basis for making judgments about whether students will achieve mandated levels of performance on such high-stakes tests. Now, perhaps more than ever before, a national premium has been placed on the importance of early identification and intervention of academic skill weaknesses. Indeed, it is quite clear that the probability of effectively remediating academic skill deficiencies increases exponentially the earlier such difficulties are detected (Juel, 1988). This, coupled with recent national legal policy initiatives calling for increased standards and documentation of progress over time (e.g., NCLB, 2001) place high importance on the early identification of academic skill deficiencies. Moreover, at the practice level, increased calls for problem-solving assessment and the use of a three-tier model of screening and monitoring of progress in response to instruction have created a need to identify reasonable and expected levels of academic performance across grades (Fuchs, Mock, Morgan, & Young, 2003; Grimes & Kurns, 2003; Kovaleski, 2003).

Stage and Jacobsen (2001), for example, examined the performance of 174 Grade 4 students who were administered R-CBM oral reading fluency passages three times during the academic year as part of benchmark assessment. Growth curve analysis was conducted

to examine the relationship between students' slope in oral reading fluency performance and the Washington Assessment of Student Learning (WASL). Results indicated a significant relationship between R-CBM slope and point estimates and the WASL. On the basis of the correlational findings, R-CBM cut scores were determined that predicted "meets standard" on the WASL. A diagnostic accuracy analysis indicated that R-CBM was able to classify correctly 74% of the participants as to whether they met or failed to meet standards on the WASL.

Similarly, Good, Simmons, and Kameenui (2001) explored a continuum of fluency-based indicators of foundational early literacy skills to predict emerging reading outcomes as well as performance on the Oregon Statewide Assessment (OSA). Significant relationships between R-CBM and performance on high-stakes assessments were observed and the utility of fluency-based benchmark goals was again supported with 96% of children who met the Grade 3 oral reading fluency benchmark goal meeting or exceeding expectations on the OSA. Likewise, McGlinchey and Hixson (2004) found a high degree of association between R-CBM and performance on the Michigan Educational Assessment Program (MEAP). R-CBM was found to be an efficient predictor of MEAP performance compared to base rates of failing and passing the MEAP. More recently, Silbergitt and Hintze (in press) examined the performance of over 2,000 students who were administered R-CBM benchmark assessments in the Spring of Grades 1, 2, and 3 and the Minnesota Comprehensive Assessment (MCA) also in the Spring of Grade 3. Results again suggested a significant relationship between R-CBM and the MCA. Not surprisingly, the relationship was strongest for those R-CBM assessments that were temporally closer to the administration of the MCA (i.e., Spring of Grade 3) than for those that were farther removed in time (i.e., Spring of Grade 1). Interestingly, R-CBM was able to predict with a high degree of accuracy (greater than 80%) those students who were likely to pass the MCA as far back as the Spring of Grade 1. That is, student oral reading fluency proficiency in

Grade 1 was significantly related to and predictive of MCA performance 2 years in advance.

Collectively, the results of these studies point to the sensitivity of R-CBM as a measure of progress over time, as well as a dynamic indicator of overall reading growth and development. As a general outcome measure, a major strength of R-CBM is in its ability to serve as a broad signal of the multifaceted construct of reading and its ability to index student performance across a variety of contexts. The purpose of this study was to replicate and extend previous research that has examined R-CBM and its relationship with high-stakes testing. Although previous research has examined the concurrent and predictive validity of R-CBM over short time durations (usually within the same year), questions still exist regarding the ability of the R-CBM oral reading fluency metric to predict reading performance over longer time durations. In particular, the purpose of the current study was twofold. First, the current study sought to compare commonly used statistical approaches to standard setting and determining cut scores (i.e., discriminant analysis, logistic regression, and receiver operator characteristic [ROC] curves). The reason for doing so was to compare the results of each analysis and come to some determination as to which one (or combination of the three) might be most advantageous for schools and practitioners to use when determining R-CBM performance standards. In addition, a second goal of the current study was to compare two different procedural models for prediction over time (i.e., constant prediction to performance on a high-stakes test or prediction to successive R-CBM benchmarking periods). The reason for doing so was again to inform practice in the hope of determining which prediction method would produce the most sensitive set of cut-score indicators for use in early identification and intervention.

## Method

### Participants

Participants for the study included 1,766 students from seven elementary schools that

were part of a regional collaborative school district in the north central U.S. The participants were drawn from five consecutive cohorts of first grade students who had complete sets of data over the period from first to third grade. Out of a total of 2,675 students, 1,815 initially qualified for participation (69% of the sample). There were no differences between the cohorts with respect to performance on the MCA. An a priori power analysis (Cohen, 1988, 1992; Hintze, 2000) was conducted suggesting that a sample size this large would provide adequate power (.80) for main effects assuming a small effect size (.10) and an alpha level of .05. Specifically the sample consisted of 49% girls and 51% boys. Ethnic breakdown was 3% Native American, 1% Asian or Pacific Islander, 1% Hispanic, 1% Black not of Hispanic Origin, and 94% White not of Hispanic Origin. Additionally, 5% of the students were receiving special educational services and 30% were eligible for free or reduced price lunch at the time of the MCA. Data were collected on five successive cohorts of students longitudinally over 3-year periods, with the first cohort beginning in the 1996-97 school year and continuing through the 1998-99 school year, and the most recent cohort beginning in the 2000-01 school year and continuing through the 2002-03 school year.

### Assessment Measures

**R-CBM oral reading fluency measures.** Standard benchmark reading assessment passages from first through third grades were used during the course of the 3-year investigation (Edformation, 2002). The benchmark passages were approximately 150 to 250 words in length and were purposively developed with controlled vocabulary and difficulty written by authors familiar with the teaching of reading and how students learn to read across a variety of types of literature. Table 1 provides the technical features of the benchmark reading assessment passages. As can be seen, the benchmark reading assessment passages evidenced adequate alternate form reliability, were of relatively equal difficulty within grade, and increased in difficulty developmentally across grades (Howe & Shinn, 2002).

**Minnesota Comprehensive Assessment (MCA).** The Grade 3 reading MCA is an untimed test, administered over the course of 2 school days. Students are asked to read up to eight narrative reading passages and respond to 56 multiple-choice and 4 constructed response items. Narrative passages include both fictional and nonfiction text, ranging in length from 250 to 800 words. All passages have been tested for readability using Degrees of Reading Power (DRP), with DRP scores ranging from 40-56 with an average of 48 for the test.<sup>1</sup> Content specifications of the reading MCA include: (a) identify main ideas and some supporting details within a text, (b) retell main events or ideas from a text in sequence, (c) demonstrate appropriate techniques for learning new vocabulary, (d) interpret presentations of data, (e) understand ideas not explicitly stated in a passage, (f) make predictions based on information in written material, (g) draw conclusions based on information in written material, (h) compare and contrast elements of a story or selection, (i) distinguish fact from opinion, and (j) summarize ideas and identify tone in persuasive, fictional, and documentary presentations (Minnesota Department of Education, 2003).

Student performance on the reading MCA is evaluated using a 5-point proficiency scoring rubric. Lower levels of the scoring rubric (i.e., Levels I, II, and III) include an assortment of skill development ranging from student gaps in knowledge and skills necessary for satisfactory work to increasing proficiency with grade level material. The upper end of the scoring rubric (i.e., Levels IV and V) characterize student performance as working above grade level and proficient with challenging subject matter to superior performance, well beyond what is expected at the grade level. Students need to obtain a proficiency level of at least III (or a standard score of 1420) on the MCA to be considered proficient.

### Procedures

Participants were assessed eight times with R-CBM measures beginning in the winter of Grade 1 and continuing each Fall, Winter, and Spring until the Spring of Grade 3. R-

**Table 1**  
**Technical Features of the Standard Benchmark Reading**  
**Assessment Passages**

Grade	Passage	Alternate Form Reliability	Lexile
1	1	.91	240
	2	.91	210
	3	.89	250
2	1	.81	420
	2	.80	440
	3	.85	470
3	1	.85	630
	2	.83	560
	3	.87	570

*Note.* Alternate form reliability is equal to the mean correlation for each alternate-form standard benchmark reading assessment passage. The lexile measure is a specific number assigned to any text indicating the reading demand of the text in terms of the semantic difficulty (vocabulary) and syntactic complexity (sentence length). A lexile unit is equivalent to 1/1000<sup>th</sup> of the difference between the comprehensibility of basal primers (the midpoint of first grade text) and the comprehensibility of an electronic encyclopedia (the midpoint of workplace text). The lexile scale ranges from 200 (beginning readers) to 1700 (advanced text) (see the Lexile Framework for Reading for a full description).

CBM measures were administered by trained staff (e.g., general and special education teachers, reading specialists, school psychologists) in accordance with standard R-CBM administration and scoring procedures (Shinn & Shinn, 2002). Likewise, students were administered the reading portion of the MCA in the spring of third grade. Scores for each of the measures were obtained for all students. That is, no students were missing any R-CBM or MCA information during the course of the 3-year assessment period.

## Results

### Data Analytic Plan

Analysis of student data was conducted in three steps. First, data were screened for outliers, distributional properties, and parametric assumptions. In addition, descriptive statistics including correlational findings were inspected. The second and third steps included an analysis of R-CBM cut scores comparing three statistical procedures: (a) discriminant

analysis, (b) logistic regression, and (c) receiver operator characteristic (ROC) curves. These three statistical procedures were chosen because they represent common approaches to establishing cut scores for diagnostic measures, and each procedure has certain advantages depending on the research question in comparison to the others. For example, ROC allows users to model a number of different cut scores across a variety of assessment situations. As such, one could develop a cut score for “screening” decisions, another for “classification” decisions, and others, all in one analysis. The potential drawback, however, is that the choice of cut-score and the definition of the assessment situation are subjective to the researcher. Discriminant analysis and logistic regression both maximize correct classifications using a statistical model. The difference between the two is that discriminant analysis tries to maximize both true positives (i.e., in this study those who are likely to fail the MCA) and true negatives (i.e., those likely to pass the MCA). In trying to balance across the two de-

cisions some inaccurate classifications are inevitable. Logistic regression attempts to maximize only true positives (or maximize those who are likely to fail the MCA). This is done, however, at the expense of false classifications as well. Importantly, different cut scores will be produced by each as a function of the research question. The extent of these differences may have implications for practice as school-based professionals use R-CBM to predict performance on high-stakes tests.

In the first analysis of these statistical comparisons, student R-CBM scores at each benchmarking period (i.e., winter Grade 1, spring Grade 1, fall Grade 2, winter Grade 2, spring Grade 2, fall Grade 3, winter Grade 3, spring Grade 3) served as the predictors (i.e., independent variable) and student reading MCA scores served as the criterion (i.e., dependent variable). In this case, reading MCA scores were dichotomized so that scores of 1420 and above were considered “passing” and scores below 1420 “failing.”<sup>2</sup> R-CBM cut scores for each benchmarking period were then determined using student reading MCA performance as the criterion standard. In the second analysis, a cut-score for R-CBM spring Grade 3 performance was initially determined using reading MCA performance as the criterion standard. Once set, the winter Grade 3 cut score was determined using the spring Grade 3 cut score as the criterion standard (i.e., dependent variable). This process continued in a sequential rearward manner with each R-CBM benchmark serving as the criterion for the immediate previous benchmark (e.g., R-CBM spring of Grade 3 serving as the criterion for winter of Grade 3; winter of Grade 3 serving as the criterion for the fall of Grade 3; fall of Grade 3 serving as the criterion for the spring of Grade 2; and so on). The rationale for these latter analyses was to (a) compare the cut scores provided by the three different statistical approaches, and (b) compare these findings across two different criteria (i.e., using a fixed criterion as in the MCA or a criterion that changes in response to developmental changes as in the benchmarking process).

## Data Screening

Prior to analysis, MCA and R-CBM scores were examined for accuracy of data entry, missing values, outliers, and the fit between their distributions and assumptions of multivariate analysis. Results of the data screening indicated the presence of 49 multivariate outliers as determined through Mahalanobis distance ( $p < .001$ ). All 49 outliers were deleted, leaving 1,766 cases for analysis. Table 2 presents the descriptive statistics for all assessment measures. As can be seen, on average participants scored above the cut score of 1420 on the reading portion of the MCA ( $M = 1,471$ ) with approximately 65% of the sample “passing” or meeting proficiency standards. In addition, R-CBM growth within and across grades was quite stable, illustrating the developmental nature of this growth metric.

Evaluation of distributional properties suggested that at the earlier R-CBM benchmarks (i.e., Grade 1 and fall of Grade 2) aggregated student performance was positively skewed and leptokurtic. This, however, is not surprising given the fact that students’ reading fluency skills at these time periods are still emerging and it is likely the case that most students perform similarly with respect to this skill (Adams, 1990; National Reading Panel, 2000). Otherwise, multivariate assumptions of linearity, homoscedasticity, multicollinearity, and singularity appeared suitable for further analyses.

## Analysis of Predictive Validity

Table 3 presents the degree of association among the MCA and R-CBM variables. As can be seen, the predictive validity of R-CBM to the MCA was significant at all time periods. Not surprisingly, R-CBM was more strongly correlated with the MCA when the two assessments were collected in closer proximity as compared to farther apart in time. In addition, the R-CBM measures were strongly related to each other, with those measures collected within a particular grade level more highly correlated than measures across grade levels. Results of this analysis suggest that R-CBM has strong validity in predicting MCA

**Table 2**  
**Descriptive Statistics of the Sample ( $N = 1,766$ )**

Assessment Measure	<i>M</i>	( <i>SD</i> )	High	Low
MCA	1471	(193)	790	2120
R-CBM Spring Grade 3	118	(42)	5	298
R-CBM Winter Grade 3	103	(40)	0	307
R-CBM Fall Grade 3	78	(37)	0	240
R-CBM Spring Grade 2	97	(38)	2	252
R-CBM Winter Grade 2	80	(37)	1	243
R-CBM Fall Grade 2	52	(32)	1	219
R-CBM Spring Grade 1	59	(33)	1	211
R-CBM Winter Grade 1	30	(26)	0	203

performance and demonstrates itself as a strong construct of reading.

#### **Analysis of the Diagnostic Accuracy of R-CBM**

Tables 4 through 6 present the results of the diagnostic accuracy analyses for the three statistical methods using the two different criterion measures (i.e., MCA and R-CBM benchmark assessments). Here, diagnostic accuracy refers to the ability of an instrument to distinguish between two diagnostic alternatives and to select the one that is correct (Swets, Dawes, & Monahan, 2000). The top half of Tables 4 through 6 present the diagnostic accuracy statistics where the predictive measures are represented by the R-CBM benchmark assessments at each grade level, and the MCA as the criterion measure. For example, W-G1 : MCA refers to the winter of Grade 1 R-CBM benchmark assessment as the predictive measure, and the MCA as the criterion measure. The bottom half of the tables again presents the R-CBM measures as the predictive measures, but the criterion measures are represented by the next successive benchmark assessment. For example, W-G1 : S-G1 refers to the winter of Grade 1 R-CBM benchmark assessment as the predictive measure and the spring of Grade 1 R-CBM benchmark assess-

ment as the criterion measure. The diagnostic accuracy statistics (i.e., Sensitivity, Specificity, PPP, and NPP) represent the probability of making a correct decision for each type of diagnostic accuracy analysis.

Most typically, diagnostic accuracy is addressed by way of a conditional probability analysis. Conditional probability refers to the likelihood of selected diagnostic outcomes, assuming that a true diagnostic status is known. There are four possible outcome proportions that result from a diagnostic accuracy analysis: (a) sensitivity, (b) specificity, (c) positive predictive power (PPP), and (d) negative predictive power (NPP). Sensitivity and specificity both refer to the proportion of agreement between the predictor and criterion measures, or in other words the accuracy of the predictor measure to identify the presence or absence of a given condition. Positive and negative predictive power, which are measures of efficiency, both refer to the probability that a predictor measure will correctly discriminate between who will be identified or not by the criterion measure, once a diagnostic status is known (Tatano-Beck & Gable, 2001). More specifically:

- *Sensitivity* (i.e., true-positive rate) refers to the probability that when a diagnostic status is present on the criterion, the indi-

**Table 3**  
**Correlations Among Assessment Measures (N = 1,766)**

	MCA	S-G3	W-G3	F-G3	S-G2	W-G2	F-G2	S-G1
MCA								
S-G3	.69							
W-G3	.68	.94						
F-G3	.66	.90	.93					
S-G2	.68	.91	.93	.93				
W-G2	.68	.88	.91	.91	.95			
F-G2	.61	.81	.85	.89	.88	.92		
S-G1	.58	.77	.81	.82	.85	.90	.91	
W-G1	.49	.63	.67	.70	.70	.76	.83	.87

*Note.* S-G3 = Spring of Grade 3; W-G3 = Winter of Grade 3; F-G3 = Fall of Grade 3; S-G2 = Spring of Grade 2; W-G2 = Winter of Grade 2; F-G2 = Fall of Grade 2; S-G1 = Spring of Grade 1; W-G1 = Winter of Grade 1. All correlations significant at the  $p < .01$  level.

**Table 4**  
**Diagnostic Accuracy Statistics for Discriminant Analysis**

Predictor : Criterion	Cut Score	Sensitivity	Specificity	PPP	NPP
S-G3 : MCA	109	.65	.87	.79	.52
W-G3 : MCA	95	.63	.87	.81	.74
F-G3 : MCA	71	.63	.87	.81	.74
S-G2 : MCA	90	.64	.87	.79	.75
W-G2 : MCA	72	.64	.86	.79	.75
F-G2 : MCA	46	.60	.86	.79	.71
S-G1 : MCA	54	.56	.86	.81	.65
W-G1 : MCA	27	.50	.87	.86	.52
W-G3 : S-G3	112	.95	.77	.87	.93
F-G3 : W-G3	83	.94	.81	.87	.90
S-G2 : F-G3	100	.92	.85	.88	.90
W-G2 : S-G2	80	.92	.87	.89	.91
F-G2 : W-G2	52	.87	.91	.92	.85
S-G1 : F-G2	61	.86	.89	.92	.81
W-G1 : S-G1	32	.82	.93	.97	.69

*Note.* S-G3 = Spring of Grade 3; W-G3 = Winter of Grade 3; F-G3 = Fall of Grade 3; S-G2 = Spring of Grade 2; W-G2 = Winter of Grade 2; F-G2 = Fall of Grade 2; S-G1 = Spring of Grade 1; W-G1 = Winter of Grade 1.

vidual will be identified positively by the predictor (i.e., the probability that those who did not pass the MCA would have been predicted to fail on the basis of their R-CBM score).

- *Specificity* (i.e., true-negative rate) refers to the probability that when a diagnostic status is absent on the criterion, the individual will not be identified by the predictor (i.e., the probability that those who did pass the MCA would have been predicted to pass on the basis of their R-CBM score).
- *Positive predictive power* (PPP) refers to the likelihood that an individual who scores below the cut score on the predictor measure will in fact have the condition of interest, based on the outcome of the criterion measure (i.e., the probab-

ity that those who were predicted to fail the MCA on the basis of their R-CBM score did in fact fail the MCA).

- *Negative predictive power* (NPP) is the likelihood that an individual who scores above the cut score on the predictor actually does not have the condition based on the criterion score (i.e., the probability that those who were predicted to pass the MCA on the basis of their R-CBM score did in fact pass the MCA).

Results of the *discriminant analysis* indicate that using successive R-CBM benchmark assessments as the criterion measure results in consistently higher cut scores as compared to using the MCA as the criterion (see Table 4). These findings suggest that the cut scores derived using R-CBM in a successive fashion across grades to ultimately predict per-

**Table 5**  
**Diagnostic Accuracy Statistics for the Logistic Regression Analysis**

Predictor: Criterion	Cut Score	Sensitivity	Specificity	PPP	NPP
S-G3 : MCA	96	.75	.82	.66	.88
W-G3 : MCA	82	.73	.82	.65	.87
F-G3 : MCA	58	.72	.82	.66	.86
S-G2 : MCA	77	.75	.82	.65	.88
W-G2 : MCA	60	.76	.82	.65	.88
F-G2 : MCA	34	.69	.81	.66	.83
S-G1 : MCA	40	.69	.80	.62	.84
W-G1 : MCA	16	.65	.78	.58	.83
W-G3 : S-G3	81	.85	.93	.84	.93
F-G3 : W-G3	56	.85	.93	.83	.93
S-G2 : F-G3	75	.87	.93	.83	.95
W-G2 : S-G2	55	.90	.95	.88	.96
F-G2 : W-G2	28	.86	.95	.86	.95
S-G1 : F-G2	36	.80	.93	.80	.92
W-G1 : S-G1	14	.79	.91	.75	.93

*Note.* S-G3 = Spring of Grade 3; W-G3 = Winter of Grade 3; F-G3 = Fall of Grade 3; S-G2 = Spring of Grade 2; W-G2 = Winter of Grade 2; F-G2 = Fall of Grade 2; S-G1 = Spring of Grade 1; W-G1 = Winter of Grade 1.

formance on the MCA lead to improved precision in identifying those students who were likely to fail the MCA as compared to the consistent use of the MCA as the criterion. Indeed, estimations of sensitivity using the MCA as the consistent criterion resulted in inordinately low levels of precision of prediction (all .65 or lower). Furthermore, using successive R-CBM benchmarks to predict MCA performance appeared more efficient as evidenced by consistently higher PPP and NPP. Both approaches worked about equally well with respect to identifying those students who were likely to pass the MCA (i.e., specificity).

Results of the *logistic regression* analysis predicted roughly equal cut scores across the two different criterion measures (i.e., MCA and R-CBM benchmark assessments). As was the case of discriminant analysis, using R-CBM

benchmark assessments in a successive fashion as the criterion resulted in higher levels of sensitivity, PPP, and NPP (see Table 5). In addition, R-CBM benchmark assessment also led to higher levels of specificity (i.e., ability to predict who would pass the MCA). Interestingly, although the two criterion methods differed with respect to diagnostic accuracy, the predicted cut-scores were highly similar.

Finally, the use of *ROC curves* indicated that using R-CBM benchmark assessments as the criterion resulted in consistently lower cut scores as compared to using the MCA. This is in direct contrast to discriminant analysis which found R-CBM to result in consistently higher cut scores as compared to the MCA. Moreover, as were the results with logistic regression, R-CBM resulted in higher levels of sensitivity, specificity, PPP, and NPP (see Table 6). Over-

**Table 6**  
**Diagnostic Accuracy Statistics for the ROC Curve Analysis**

Predictor: Criterion	Cut Score	Sensitivity	Specificity	PPP	NPP
S-G3 : MCA	109	.79	.76	.65	.87
W-G3 : MCA	93	.74	.83	.70	.85
F-G3 : MCA	68	.70	.86	.73	.84
S-G2 : MCA	88	.67	.87	.74	.82
W-G2 : MCA	71	.65	.88	.75	.82
F-G2 : MCA	41	.62	.89	.76	.81
S-G1: MCA	46	.58	.90	.77	.79
W-G1 : MCA	20	.54	.92	.79	.78
W-G3 : S-G3	76	.88	.94	.62	.99
F-G3 : W-G3	49	.88	.95	.69	.99
S-G2 : F-G3	65	.88	.95	.76	.99
W-G2 : S-G2	43	.89	.97	.80	.99
F-G2 : W-G2	22	.89	.99	.90	.99
S-G1 : F-G2	28	.89	.99	.99	.99
W-G1 : S-G1	12	.80	.99	.99	.98

*Note.* S-G3 = Spring of Grade 3; W-G3 = Winter of Grade 3; F-G3 = Fall of Grade 3; S-G2 = Spring of Grade 2; W-G2 = Winter of Grade 2; F-G2 = Fall of Grade 2; S-G1 = Spring of Grade 1; W-G1 = Winter of Grade 1.

all, using R-CBM in a successive manner across grades to ultimately predict performance on the MCA appeared to be a more accurate and efficient approach as compared to the consistent use of the MCA as the criterion.

### Discussion

The purpose of this study was to compare different statistical and methodological approaches to standard setting and determining cut scores using R-CBM and performance on high-stakes tests. Overall, the findings support and extend previous work in the area and support the use of R-CBM as a powerful predictor of more global measures of reading (Good, Simmons, & Kameenui, 2001; McGlinchey & Hixson, 2004; Silbergitt & Hintze, in press; Stage & Jacobsen, 2001). In

particular, results of this study point to a number of interesting findings. First, each statistical procedure was able to set cut scores that yielded adequate levels of both diagnostic accuracy and efficiency. As such, R-CBM appears to be an efficient method for predicting performance on high-stakes tests demonstrating the ability to predict those students who are likely to pass reading portions of such tests as far back as first grade. Evaluating the three statistical approaches produces generally consistent findings. For the most part, each approach produced cut scores that yielded higher levels of specificity and negative predictive power as compared to sensitivity and positive predictive power, respectively. These findings are consistent with the results of Silbergitt and Hintze (in press) and are not surprising given

that each method attempted to maximize the number of true negatives, students who would be correctly predicted to ultimately pass the MCA. In the case of discriminant analysis, students are classified into groups using posterior probabilities. By examining a set of variables describing a population, discriminant analysis determines the probability of membership in discrete groups (Huberty, 1994; Klecka, 1980). Comparatively, logistic regression uses maximum likelihood estimation, whereby group membership is calculated as a probability (0 to 100%) that a person is a member of a group (Neter, Kutner, Nachtsheim, & Wasserman, 1996). Finally, ROC curves plot the sensitivity and specificity of a predictor for all possible values of the cut score (Swets, 1996). The fact that all three approaches produced consistent diagnostic accuracy results bodes well for R-CBM as a predictor of MCA performance.

A second interesting finding of the study was that using R-CBM to set cut scores in a successive manner from one benchmarking period to the next across grades appeared to be a more accurate and efficient method than using a high-stakes test consistently as the criterion regardless of the grade level. Again, this is not surprising because predictions made from one benchmark period to the next occur more closely in time to each other than to the MCA, which, depending on the benchmark period in question, can be far removed in time. If this is in fact the case, setting cut scores to what in essence is a moving target (i.e., the changing nature of the benchmarks) might prove problematic at both practical and policy levels. Changes in the student composition of a school district, curricula changes, instructional changes, etc., all may influence the nature of student reading and thus benchmark levels of reading. In cases such as these attempting to determine cut scores in a successive manner may prove problematic.

For practitioners and policy makers who might want to determine R-CBM cut scores to predict performance on high-stakes tests the question then is one of “which statistical approach and prediction method makes sense?” The answer to this question is certainly not easy

and rests largely on the resources of the school district both in terms of expertise and data collection abilities. For districts with some level of expertise, ROC curves provide a highly flexible means for determining cut scores across a wide variety of assessment decisions. For example, different cut scores could be developed for “screening” decisions, another set for “classification” decisions, and yet another for “entitlement” decisions. In this sense, ROC curves allow the users to model the level of diagnostic accuracy that can be expected across a variety of cut scores while still allowing for maximal diagnostic accuracy and efficiency.

For districts that prefer one set of scores for classification purposes it would seem that either discriminant analysis or logistic regression would be suitable alternatives. Interestingly, results of this study suggest that logistic regression produces cut scores that are (a) both accurate and efficient, and (b) highly similar regardless of whether the criterion is a fixed standard (as in the case of the MCA) or one that moves as a function of developmental changes in the student population (as in the case of the R-CBM benchmark assessments). Whereas discriminant analysis yielded consistently higher cut scores for the benchmark assessment as compared to the MCA condition (range of 5 to 17 points different with an overall average difference of 9 points), logistic regression produced a consistent set of cut scores across both criterion conditions (range of 1 to 6 points different with an overall average difference of 3 points). In this case it would appear that the most parsimonious solution to the problem would be to use logistic regression with R-CBM serving as the predictor and the high-stakes test as the criterion.

In addition to standard setting and setting cut scores, results of the current study have implications for practitioners within the context of early identification, primary prevention, and response to intervention within a three-tiered model of service delivery (Grimes & Kurns, 2003; Kovaleski, 2003). Indeed, in addition to determining risk status for the probability of failing high-stakes tests, the current cut scores can easily be used for identifying those students who are currently

at risk for developing reading problems or who are in need of compensatory support services in reading.

Findings of the current study underscore the utility of using these results as part of a school wide screening effort in hopes of staying off the deleterious effects of unidentified and prolonged skill deficits in reading (Juel, 1988). Moreover, growth approximations may be determined for those students whose reading development is being progress monitored. By comparing current levels of student performance to desired levels, practitioners are able to determine goal level growth estimates for use in progress monitoring. For example, if a third grade student were currently reading 60 words correct per minute in third grade material in the fall and the desired level of performance was 110 words correct per minute by the end of the school year (e.g., 25 weeks), dividing the difference between current and desired performance by the number of weeks left in the school year would provide growth targets for progress monitoring (i.e., a goal of 2 words growth per week). The use of cut scores in this manner allows practitioners to make decisions regarding the relative standing of an individual (e.g., how does a student's score compare to others in his or her grade), as well as evaluating within individual progress (e.g., how has a student's skills improved over time relative to where he or she started).

### Limitations

Although the results of the current study are certainly encouraging for those who would like to use CBM measures to predict student performance on high-stakes tests, the results must nevertheless be interpreted in light of a number of possible limitations (Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2002). First, the effects that the changes in instrumentation would have on the observed results are unknown. Whether similar findings would be evidenced with different examiners, different R-CBM measures, different state tests, etc., is hard to determine. Follow-up studies to this one are likely to employ bootstrapping methods in an attempt to form confidence in-

tervals for the cut scores to account for such variations in assessment arrangements. Second, the extent to which the current findings generalize to other student samples is unknown. Although the sample size of the current study was indeed robust, the generalizability of the findings need to be determined in replicate studies.

### Summary and Conclusions

Overall, results of the current study suggest that R-CBM is strongly associated with MCA performance at each grade level and is both accurate and efficient in predicting those students who are likely to pass the reading portion of the MCA. Using a variety of statistical approaches and predictive methods, practitioners can determine cut scores that will enable them to effectively screen for those students—as early as first grade—who are at-risk for failing high-stakes state testing programs. When infused within a primary preventative program of identification and intervention, the use of R-CBM in this manner will provide value added to any school-based assessment model.

### Footnotes

<sup>1</sup> The DRP is a measure of text difficulty based on what children are able to read considering both semantic and syntactic difficulty. Average text difficulty ranges from 31 to 76. For example, primary school textbooks (e.g., "Clifford the Big Red Dog," "Frog and Toad are Friends") have an average DRP score of 40 whereas the front page of newspapers have an average DRP score of 70.

<sup>2</sup> Scores of 1,420 and above are considered to have met proficiency standards (i.e., Levels III, IV, and V) according to Minnesota state standards.

### References

- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand-McNally.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52, 219-232.

- Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*, 184-192.
- Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36-45.
- Deno, S. L., Reschly-Anderson, A., Lembke, E., Zorka, H., & Callender, S. (2002, March). *A model for school wide implementation: A case example*. Paper presented at the National Association of School Psychologists, Chicago, IL.
- Edformation. (2002). *AIMSweb standard benchmark reading assessment* passages. Eden Prairie, MN: Author.
- Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice, 18*, 157-171.
- Fuchs, D., Roberts, P. H., Fuchs, L. S., & Bowers, J. (1996). Reintegrating students with learning disabilities into the mainstream: A two-year study. *Learning Disabilities Research & Practice, 11*, 214-229.
- Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*, 488-500.
- Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*, 449-460.
- Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice, 13*, 204-219.
- Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1993). Technological advances linking the assessment of students' academic proficiency to instructional planning. *Journal of Special Education Technology, 12*, 49-62.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal, 28*, 617-641.
- Good, R. H., Simmons, D. C., & Kameenui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*, 257-288.
- Grimes, J., & Kurns, S. (2003). *An intervention-based system for addressing NCLB and IDEA expectations: A multiple tiered model to ensure every child learns*. Paper presented at the Responsiveness to Intervention Symposium sponsored by the National Research Center on Learning Disabilities, Kansas City, MO.
- Hintze, J. L. (2000). *PASS power analysis*. Kaysville, UT: NCSS Statistical Software.
- Howe, K. B., & Shinn, M. M. (2002). *Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual for describing development and technical features*. Eden Prairie, MN: Edformation.
- Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley & Sons.
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437-447.
- Klecka, W. R. (1980). *Discriminant analysis*. Beverly Hills, CA: Sage.
- Kovaleski, J. F. (2003). *The three-tier model for identifying learning disabilities: Critical program features and system issues*. Paper presented at the Responsiveness to Intervention Symposium sponsored by the National Research Center on Learning Disabilities, Kansas City, MO.
- Lexile Framework for Reading. (n.d.). Retrieved June 24, 2004 from <http://www.lexile.com/>
- Marston, D., & Magnusson, D. (1988). Curriculum-based measurement: District level implementation. In J. L. Graden, J. E. Zins, & M. J. Curtis (Eds.), *Alternative educational delivery systems: Enhancing options for all students* (pp. 137-172). Washington, DC: National Association of School Psychologists.
- Marston, D., Mirkin, P. K., & Deno, S. L. (1984). Curriculum-based measurement: An alternative to traditional screening, referral, and identification. *The Journal of Special Education, 18*, 109-117.
- McGlinchey, M. T., & Hixson, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*, 193-203.
- Minnesota Department of Education. (2003). *Minnesota Comprehensive Assessments: Grade 3 reading test specifications*. Roseville, MN: Author.
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models*. Boston, MA: McGraw-Hill.
- No Child Left Behind Act of 2001*. Pub. L. No. 107-110.
- Powell-Smith, K. A., & Stewart, L. H. (1998). The use of curriculum-based measurement on the reintegration of students with mild disabilities. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 254-307). New York: Guilford Press.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
- Shinn, M. R. (1989a). Identifying and defining academic problems: CBM screening and eligibility procedures. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 90-129). New York: Guilford Press.
- Shinn, M. R. (Ed.). (1989b). *Curriculum-based measurement: Assessing special children*. New York: Guilford Press.
- Shinn, M. R. (1995). Best practices in curriculum-based measurement and its use in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology—III* (pp. 547-567). Washington, DC: National Association of School Psychologists.
- Shinn, M. R. (2002). Best practices in using curriculum-based measurement in a problem-solving model. In A.

- Thomas & J. Grimes (Eds.), *Best practices in school psychology* (Vol. 4, pp. 671-697). Silver Spring, MD: National Association of School Psychologists.
- Shinn, M. R., & Shinn, M. M. (2002). *Administration and scoring of reading curriculum-based measurement (R-CBM) for use in general outcome measurement*. Eden Prairie, MN: Edformation.
- Silberglitt, B., & Hintze, J. M. (in press). Formative assessment using oral reading fluency cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment*.
- Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review*, 30, 407-419.
- Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Mahwah, NJ: LEA.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, 1, 1-26.
- Tatano-Beck, C., & Gable, R. K. (2001). Further validation of the postpartum depression screening scale. *Nursing Research*, 50, 155-164.
- Tilly, W. D., & Grimes, J. (1998). Curriculum-based measurement: One vehicle for systematic educational reform. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 32-88). New York: Guilford Press.

John M. Hintze, PhD, is an Associate Professor of School Psychology at the University of Massachusetts at Amherst. He received his doctorate from Lehigh University in 1994 and prior to that was a practitioner in the public schools for 10 years. His research interests are in CBM and various forms of progress monitoring, research design, and data analysis that informs practice.

Benjamin Silberglitt, PhD, received his doctorate in Educational Psychology (School Psychology) in 2003 from the University of Minnesota-Twin Cities. He works as the Outcomes Manager for the St. Croix River Education District, MN.

