

Curriculum-Based Measures and Performance on State Assessment and Standardized Tests

Reading and Math Performance in Pennsylvania

Edward S. Shapiro

Milena A. Keller

J. Gary Lutz

Lehigh University

Lana Edwards Santoro

Pacific Institutes for Research

John M. Hintze

University of Massachusetts–Amherst

General outcome measures (GOMs) provide educators with a means to evaluate student progress toward curricular objectives. Curriculum-based measurement (CBM) is one type of GOM that has a long history in the research literature with strong empirical support. With the increased emphasis on instruction linked to state standards and statewide achievement tests, the relationship between CBM and these measures has been called into question. This study examined the relationships between CBM of reading, math computation, and math concepts/applications and the statewide standardized achievement test as well as published norm-referenced achievement tests in two districts in Pennsylvania. Results showed that CBM had moderate to strong correlations with midyear assessments in reading and mathematics and both types of standardized tests across school districts. The data suggest that CBM can be one source of data that could be used to potentially identify those students likely to be successful or fail the statewide assessment measure.

Keywords: *curriculum-based measurement; CBM; high stakes tests; academic achievement*

General outcome measures (GOMs) provide educators with a means to assess student performance and use that information to guide instruction (Deno, 2003; Fuchs & Deno, 1991). GOMs are standardized and are able to assess student performance over long periods of time in a consistent manner. What makes GOMs unique assessment tools is that GOMs can function as an index of student progress through the curriculum over time, allowing teachers to more carefully chart student responsiveness to their instruction (Fuchs & Deno, 1991).

Authors' Note: Correspondence concerning this article should be addressed to Edward S. Shapiro, PhD, Director, Center for Promoting Research to Practice, Lehigh University, L-111 Iacocca Hall, 111 Research Drive, Bethlehem, PA 18015; e-mail: ed.shapiro@lehigh.edu.

Curriculum-based measurement (CBM), an example and type of GOM, has been used as an evaluative measure in education for more than two decades (Deno, 2003; Fuchs, 2004; Shinn, 1989). More specifically, CBM involves evaluating progress toward the acquisition of basic skills during instruction (Deno, Espin, & Fuchs, 2002; Fuchs & Fuchs, 1999; Shinn & Bamonto, 1998; Shinn, Shinn, Hamilton, & Clarke, 2002). CBM is a GOM that is a standardized procedure conducted repeatedly over time and provides information relevant to making decisions regarding student achievement (Deno et al., 2002; Fuchs & Deno, 1991; Fuchs & Fuchs, 1999; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Shinn et al., 2002). Also, because the assessment methods mirror skills that students use during everyday instruction such as reading aloud, CBM is considered to be an “authentic” way to assess students (Deno et al., 2002; Shepard, 2000). These procedures can be used in several academic areas including reading, math, and writing; have established technical characteristics; and have been shown to be related to overall student achievement (Deno, 1985; Deno et al., 2002; Good & Jefferson, 1998; Shinn, 1989). Finally, assessing student progress using CBM depicts academic growth because it is a classroom-based method that provides repeated samples of performance (Fuchs & Fuchs, 1999). This ultimately allows for the modeling of change over time, which is invaluable to teachers tracking the progress of their students and making instructional modifications (Fuchs & Fuchs, 1999).

Research surrounding CBM as a general outcome measure has examined various uses of CBM data and the means by which CBM can influence instruction and learning in general and special education (Deno et al., 2002). Although the use of CBM apparently has numerous advantages, questions have been raised about the relationship of CBM to standardized assessments that are used as achievement accountability measures by states. Given the increased national emphasis on standardized testing and the development of curriculum standards, this criticism is an issue central to validity of CBM. Most states, as well as many school districts, have requirements to conduct and disseminate the results of standardized tests to indicate the effectiveness of their educational efforts (Shinn et al., 2002; Thurlow & Thompson, 1999). In fact, the passage of the No Child Left Behind (NCLB; 2000) law requires that all states use some form of statewide achievement test to determine district accountability for student progress. The standardized tests being used by school districts to demonstrate student performance outcomes are most often based on curriculum standards established by the state (Braden, 2002; Erikson, Ysseldyke, Thurlow, & Elliott, 1998; Linn, 2000).

Curriculum standards are the result of governmental education initiatives designed to improve the nation’s schools (Braden, 2002; Linn, 2000). The initiatives urged states to develop rigorous curriculum standards and then to devise their own standards-based assessment systems (Braden, 2002; Linn, 2000). As a result of these government acts, all 50 states have curriculum standards and a means by which to measure student acquisition of the required content (Braden, 2002). Curriculum or content standards define what students should know and influence the way standards are evaluated (Braden, 2002; Erikson et al., 1998; Linn, 2000). Content standards are not the only standards receiving attention; performance standards are also of great importance (Linn, 2000). Performance standards are associated with specific grades and outline what knowledge students should be able to demonstrate at a particular level (Braden, 2002). Regardless of the type of standard, content or performance, standardized assessments of the acquisition of the material is required.

Several studies have examined the relationship between CBM and statewide standardized achievement tests, especially in reading. Powell-Smith (2004) reported on the results of stud-

ies that examined the relationships between CBM measures of reading (oral reading fluency) and outcomes on statewide assessments in Colorado (Shaw & Shaw, 2002), Florida (Buck & Torgeson, 2003; Castillo, Torgeson, Powell-Smith, & Al Otaiba, 2003), Illinois (Sibley, Biwer, & Hesch, 2001), Michigan (McGlinchey & Hixson, 2004), Minnesota (Hintze & Silbergliitt, 2005), North Carolina (Barger, 2003), Oregon (Crawford, Tindal, & Stieber, 2001; Good, Simmons, & Kame'enui, 2001), and Washington (Stage & Jacobsen, 2001). These studies found the correlation between performance on a measure of oral reading fluency taken at the end of third or fourth grade and performance on the state assessments to range between .44 (Washington) and .79 (Illinois). On average, most studies reported correlations in the .60 to .75 range. Considering the range of states, differences in types of measures, and the fact that each state developed its own assessment tool based on its own curriculum standards, the remarkable consistency of the relationship between student performance on a 1-minute oral reading passage and the high-stakes tests was quite powerful in suggesting the link between CBM and the statewide assessment measure.

Few studies have reported outcomes of relationships between CBM and statewide assessments in math. Helwig, Anderson, and Tindal (2002) examined the effectiveness of a CBM math concept task at predicting eighth-grade student scores on a computer adaptive test of math achievement designed to approximate a state (Oregon) standardized math achievement measure. All students were presented with a CBM math probe containing 48 items including both problem-solving and computation tasks and were not given a time limit. The computer adaptive math assessment was used in place of the actual standardized test administered in the spring to all students because the results of the actual test were not available at the time of the study (Helwig et al., 2002). The computer adaptive test was provided to the researchers by the Department of Education (DOE) and included items similar to those of the statewide math exam and generated scores on the same scale. Both the DOE and the software developer considered the computerized test to be a valid substitute for the original standardized measure. The CBM task and the computerized test were administered within 2 to 3 weeks of each other during the spring (Helwig et al., 2002).

Results indicated that the math CBM task used in this study was effective at predicting scores on the computer adaptive test of math assessment for students in general education (Helwig et al., 2002). In fact, when the data were analyzed using discriminant function analysis, CBM math probes predicted with 87% accuracy the students who would meet the state math standards (Helwig et al., 2002). Helwig et al. (2002) noted that tasks such as CBM that can accurately estimate progress toward statewide goals in addition to monitoring classroom progress have considerable utility in planning instruction.

Although the relationship between GOMs in reading, published norm-referenced tests, and statewide assessments appears to be established across many states, the idiosyncratic nature of each assessment calls into question whether the outcomes for any single state will apply across others. In addition, there are limited findings reported for outcomes of relationships between CBM and statewide assessments in mathematics.

This study adds to the existing studies examining the relationship between CBM and standardized assessments including both state assessments and published norm-referenced standardized achievement tests (Stanford Achievement Test–Ninth Edition [SAT-9; Harcourt Brace Educational Measurement, 1996] and Metropolitan Achievement Test–Eighth Edition [MAT-8; Harcourt Brace Educational Measurement, 2002]) in two districts in Pennsylvania.

Table 1
Demographic Characteristics of the
Two School Districts Used in the Normative Study

District 1	District 2
Moderate-sized urban/suburban ($N = 14,442$)	Small suburban ($N = 6,851$)
14 elementary buildings	8 elementary buildings
32.8% low income	6.3% low income
8.4% limited English proficient	<1% limited English proficient
11.4% students have IEPs	10.2% students have IEPs
Average 5th-grade PSSA	Average 5th-grade PSSA
Reading = 1350	Reading = 1430
Math = 1380	Math = 1440
% students proficient on PSSA (5th grade)	% students proficient on PSSA (5th grade)
Reading = 36%	Reading = 46%
Mathematics = 28%	Mathematics = 27%
Standardized tests administered	Standardized tests administered
Grade 4 = MAT-8 Reading	Grades 2 and 4 = SAT-9 Reading and Math
Grade 5 = SDRT	

Note: IEP = Individualized Education Plans; PSSA = Pennsylvania System of School Assessment; MAT-8 = Metropolitan Achievement Test–Eighth Edition; SAT-9 = Stanford Achievement Test–Ninth Edition; SDRT = Stanford Diagnostic Reading Test.

This study also goes beyond previous research in this area by including CBM in math computation as well as math concepts/applications.

Method

Participants and Setting

Participants were taken from curriculum-based norming projects in two school districts in eastern Pennsylvania. Table 1 provides descriptive information about each of the districts, one a moderate-sized district with a mix of urban and suburban schools (District 1), the other a small suburban district (District 2). Normative samples for District 1 were collected for both reading and math during the 2002-2003 school year and consisted of a total of 1,461 and 1,477 students, respectively. The participants were drawn as a stratified random sample across six elementary schools in the district. The sample was stratified on a socioeconomic basis by selecting schools whose percentages of free-and-reduced lunch participants were less than 10%, between 10% and 35%, and greater than 35% such that the final sample was representative of the district's overall free-and-reduced lunch level of 32%. The normative sample of District 2 were collected during the 2001-2002 school year and consisted of a total of 782 students drawn as a stratified random sample across all elementary schools of the district such that the final sample was proportional to number of students in each grade and in each elementary school in the district. Students excluded from both district samples were those with active Individualized Education Plans (IEPs) other than those in Gifted and Speech/Language programs. Only students who maintained participation in the norming project throughout the entire school year served as participants for the current study. The final sample of reading data for analysis in the current study in District 1 comprised 617 stu-

dents: 191 from third grade, 213 from fourth grade, and 213 from fifth grade. The final sample of math data for analysis in District 1 comprised of 475 students: 205 from third grade, 138 from fourth grade, and 132 from fifth grade. The final sample for reading and math for analysis in District 2 comprised 431 students: 132 from third grade, 133 from fourth grade, and 166 from fifth grade.

Measures

To examine the relationship between progress monitoring and standardized assessments, two types of measures were used: (a) curriculum-based measures (i.e., probes for reading, math computation, and concepts/applications) and (b) standardized assessments (i.e., Pennsylvania System of School Assessment (PSSA; Pennsylvania Department of Education [PDE], 2002); SAT-9 (Harcourt Brace Educational Measurement, 1996), MAT-8 (Harcourt Brace Educational Measurement, 2000); Stanford Diagnostic Reading Test (SDRT; Karlson & Gardner, 1995).

Curriculum-based measures. For District 1, reading passages developed by AIMSweb (Edformation, 2005) were used. These passages are grade-based narrative reading passages of 150 to 300 words. Each passage is evaluated for readability using both the Fry (1968) formula and Lexile[®] formula. Probes were randomly selected at each grade level from the available set with the number of words read correctly per minute serving as the dependent measure. The blackline masters from Monitoring Basic Skills Progress (MBSP)–Math Computation (Fuchs, Hamlett, & Fuchs, 1998) were used to assess student progress in math computation. Each probe consisted of a single sheet of 25 mixed operation problems that are designed to assess mastery of computational skills typical for each grade level. For example, problems for Grades 1 and 2 consisted of addition and subtraction problems with and without regrouping. Simple multiplication and division facts were introduced in Grade 3 probes, with fractions and multidigit multiplication as well as simple division introduced in Grade 4 probes. Decimals, complex fractions, and multidigit division with remainders were introduced in Grade 5 and 6 probes. Each computation problem was scored by counting the number of digits correct in the final answer. The total number of digits correct was used as the dependent measure to reflect student performance on math computation. From each grade level, a set of probes were randomly selected. The blackline masters from MBSP–Math Concepts and Applications (Fuchs, Hamlett, & Fuchs, 1999) were used to assess student progress in math concepts/applications. The probes consisted of 18 (second-grade level) or 24 problems (third- to sixth-grade levels) designed to assess mastery of concepts and application of mathematical skills. The measures specifically cover the areas of counting, number concepts, names of numbers, measurement, charts and graphs, money, fractions, applied computation, and word problems. Problems began at Grade 2 and continued to increase in difficulty through Grade 6. Problems required between one and three responses and varied in type (fill-in-the-blank, multiple choice). Each part of a problem answered correctly was awarded a point and the total number of points across the probe was used as the dependent measure for math concepts/application. From each grade level, probes were randomly selected for use in the norming project.

For District 2, reading passages and multiple-skill math computation probes were developed by the school district as part of the local norming project. With respect to reading, the district had available a set of generic probes currently in use. The researchers developing the

local norms selected probes randomly from those provided and examined them to determine readability using the Fry formula (Fry, 1968). Readability results within the designated grade level of the probe were considered acceptable. Passages that did not meet this criterion were discarded and another was randomly selected and evaluated for readability. The selected passages were then shared with the district and used if deemed acceptable. The math probes were also generated by the district and were based on an examination of the scope and sequence of computational objectives for each grade. The researchers randomly selected probes from those furnished by the district and were used after it was determined that the probes were acceptable by the district personnel. The dependent measures for the CBM reading and math probes were oral reading fluency (words correct per minute [wcpm]) and digits correct per minute [dcpm], respectively).

Research using CBMs have established the technical characteristics of these assessment measures (Shinn, 1989). The reliability of CBM has been established for over a decade (Deno, 2003; Shinn, 1989). Test-retest reliability coefficients for reading CBM probes ranged from .82 to .97, interrater reliability was .99, and the reliability coefficients for parallel forms ranged from .84 to .96 (Marston, 1989). Math CBM probes have also been proven reliable with a mean reliability of .91 (Marston, 1989). For an extensive list of CBM reliability and validity studies, the reader is directed to reviews by Shinn (1989; Shinn & Bamonto, 1998).

PSSA. The PSSA examines both reading and math skills in an effort to provide information regarding student achievement (PDE, 2003). The reading assessment portion of the PSSA covers five general skill areas. Students are required to read various passages and respond to them in ways that reflect various skills including (a) learning to read independently; (b) reading critically; (c) reading, analyzing, and interpreting literature; (d) characteristics and function of the English language; and (e) research. The math portion of the PSSA examines 11 math standards ranging from basic skills such as numbers, number systems, and the relationships between numbers to more advanced skills such as trigonometry and calculus (PDE, 2003).

The PSSA returns a standard total score in reading and mathematics ranging from 700 to 2100. Passing scores have been established only for fifth grade and are as follows in reading: <1160 = below basic, 1160-1299 = basic, 1300-1479 = proficient, >1479 = advanced. In math, scores are interpreted as follows: <1170 = below basic, 1170-1299 = basic, 1300-1459 = proficient, >1459 = advanced. Scores at or above proficient are reported to be equivalent to the 43rd percentile performance on the PSSA. Scores at or above basic are equivalent to the 22nd percentile on the PSSA (Mead, Smith, & Swanlund, 2003).

The *Handbook for Report Interpretation* (PDE, 2002) outlines the most recent reliability and validity characteristics of the PSSA. With respect to evaluation of reliability, internal consistency for fifth grade students was very high with coefficients of .93 for mathematics and .90 for reading. Although there are no actual coefficients reported with respect to the validity of the PSSA, it is noted in the handbook that the primary focus was content validity. Experts in the various content areas evaluated the content validity of the PSSA including teachers and those involved in curriculum planning. These groups of experts met to construct items and develop scoring procedures and criteria and also reviewed the results of preliminary tests of the items (PDE, 2002).

SAT-9. The SAT-9 is an achievement test designed to assess student achievement in Reading (Word Study Skills, Reading Vocabulary, Reading Comprehension, and Total Reading), Math (Problem Solving, Procedures, and Total Math), Language, Spelling, and Listening (Impara & Plake, 1998). The *Mental Measurement Yearbook* (Impara & Plake, 1998) reported the reliability and validity of this assessment. Reliability indexes included the Kuder-Richardson Formula 20 (KR20) and the Kuder-Richardson Formula 21. Results for most multiple choice tests and subtests using the KR20 provided coefficients ranging from the mid .80s to .90s, and results of the KR21 were also high, ranging from the .70s to the .90s (Impara & Plake, 1998). Interrater coefficients for the writing assessments ranged from the .50s to the mid-.80s, whereas Spearman-Brown coefficients were slightly higher, ranging from .70 to the mid .90s (Impara & Plake, 1998). Validity evidence for the SAT-9 includes content and concurrent validity. Normal curve equivalents on each of the subtests as well as the total normal curve equivalent score were used as the dependent measures.

MAT-8. The MAT-8 was designed as a measure of general achievement for students from kindergarten through high school. The test at the elementary level includes assessments of reading (Sounds and Print, Reading Vocabulary, Reading Comprehension, Total Reading); mathematics (Mathematics Concepts and Problem Solving, Mathematics Computation, Total Mathematics); and language, spelling, social studies, and science as well as a total battery score. KR20 and KR21 estimates of reliability are reported in technical manuals and found to be in the .80 to .90 range. Concurrent validity with the Otis-Lennon School Ability Test and the MAT-7 were reported to be in the .60 to .85 range (Spies & Plake, 2005). Raw scores on each of the subtests were used as the dependent measures.

SDRT. The SDRT was designed to diagnose students' strengths and weaknesses in several major elements of reading (Karlsen & Gardner, 1995). The foundation for the SDRT is a developmental perspective of reading and includes four basic components: phonetic analysis, vocabulary, comprehension, and scanning (Impara & Plake, 1998). The technical properties of the SDRT include an assessment of reliability using estimates of the KR20 and KR21. The KR20 results for the four parts of the SDRT ranged from .79 to .94 (Impara & Plake, 1998). KR20 results for the total SDRT score ranged from .95 to .98 (Impara & Plake, 1998). KR21 results, according to Impara and Plake (1998), are similar, but some of the coefficients for the subtests are low. No test-retest reliability results were provided, so there is no evidence of the stability of SDRT scores over time. The alternate-form reliability, however, ranges from .86 to .88 for the total score (Impara & Plake, 1998). The total raw score on this assessment was used as the dependent measure.

Concurrent validity was assessed using correlations with the Otis-Lennon School Ability Test—Sixth Edition and previous versions of the SDRT, resulting in outcomes that are within the range expected for an analysis such as this (Impara & Plake, 1998).

Procedures

CBM data were collected prior to this study as a part of the two local norming projects. Local norms were established for both districts in reading and math computation for Grades 1 through 5 and in math concepts/applications only in District 1. Data were collected during

the school year from students in Grades 1 through 5 over a 10- to 15-day period in October, February, and May.

Results of the administration of standardized tests at different grades were used across the districts. For both districts, the PSSA was administered for third and fifth grades in the spring (March or April) of the school year in which normative data were collected. In District 1, MAT-8 data in reading and math were collected in Grade 4, and SDRT data in reading were collected in Grade 5. In District 2, SAT-9 data were collected in reading and math in Grades 2 and 4.

Results

Outcomes With PSSA

The data were examined for outliers and to ensure that the appropriate statistical assumptions were met. Descriptive statistics were used to identify cases with missing data, to examine the dispersion of the data, and to assess for normality. If cases were identified with missing data, those particular cases were eliminated from the analysis. The data were graphed as a histogram with a normal curve to assess the range and the degree to which the data were distributed normally. Through this analysis data points that were either entered incorrectly or determined to be outliers were identified and eliminated. All variables were determined to be normally distributed prior to data analysis.

Table 2 shows the Pearson product-moment correlations between the CBM reading (Oral Reading Fluency [ORF]) measure and the PSSA results obtained at fall, winter, and spring assessments across districts. All correlations were statistically significant ($p < .001$), and all except for the fall assessment for District 2 ranged between .62 and .69. Hierarchical regression analysis showed that the winter assessment period marked the strongest predictor to the PSSA scores, with spring assessments not adding significantly to explanations of variance contributing to outcomes.

Table 3 shows the outcomes for math computation across districts, and Table 4 shows the outcomes for math concepts/applications for District 1, the only district where concepts/applications data were collected. As noted in Table 3, moderate correlations between math computation and PSSA scores were found across winter and spring assessment periods, with correlations ranging from .50 to .53, all statistically significant ($p < .001$). Somewhat lower correlations were found at the fall assessment (.07 to .41), with one nonsignificant relationship (District 1, fall). A regression analysis showed the winter assessment period again as the strongest predictor to PSSA scores. Table 4 shows the outcomes for math concepts/applications in District 1 and shows moderate correlations across assessment time periods and grades. Correlations during winter and spring ranged from .56 to .64, with fall correlations of .46 and .48. All correlations were statistically significant ($p < .001$), and the regression analysis again showed the winter assessment to be the most powerful predictor of PSSA scores.

The PSSA assigns students into one of four categories based on their performance: below basic, basic, proficient, and advanced. Given that the state DOE grades districts on the basis of the percentage of those students falling below proficient, a diagnostic analysis for the fifth grade of each district for reading and math measures was conducted. At the time of this study, criteria for PSSA categories had only been established for fifth grade. The analysis was conducted using the CBM winter norm scores, as the winter assessment period was consistently

Table 2
Correlations Between CBM (ORF) in Reading and PSSA Scores Across Districts

	Grade	Fall	Winter	Spring
District 1	5 (<i>n</i> = 206)	.681	.693	.669
	3 (<i>n</i> = 185)	.647	.664	.671
District 2	5 (<i>n</i> = 127)	.245	.642	.623

Note: CBM = curriculum-based measurement; ORF = Oral Reading Fluency; PSSA = Pennsylvania System of School Assessment. All correlations were significant at $p < .001$.

Table 3
Correlations Between CBM Math Computation and PSSA Across Districts

	Grade	Fall	Winter	Spring
District 1	5 (<i>n</i> = 126)	.072	.505	.521
	3 (<i>n</i> = 190)	.408	.525	.519
District 2	5 (<i>n</i> = 119)	.250	.522	.538

Note: CBM = curriculum-based measurement; PSSA = Pennsylvania System of School Assessment. All correlations were significant at $p < .001$ except .072 (n.s.).

Table 4
Correlations Between CBM Math Concepts/Applications and PSSA for District 1

	Grade	Fall	Winter	Spring
District 1	5 (<i>n</i> = 126)	.479	.641	.561
	3 (<i>n</i> = 190)	.457	.613	.644

Note: CBM = curriculum-based measurement; PSSA = Pennsylvania System of School Assessment. All correlations were significant at $p < .001$.

found to have the strongest relationship with PSSA outcomes. Diagnostic accuracy is represented using the following descriptive statistics (Swets, Dawes, & Monahan, 2000): (a) *sensitivity* refers to the probability that the CBM score will accurately identify those students who were not successful on the PSSA; (b) *specificity* refers to the probability that the CBM score will accurately identify those students who have been successful on the PSSA; (c) *false negatives* refers to the probability that the CBM measure will fail to accurately identify students who have failed the PSSA; (d) *false positives* refers to the probability that the CBM measure will fail to accurately identify students who were successful on the PSSA; (e) *positive predictive power* refers to the probability that those students identified as failing on the CBM measure will be identified as failing on the PSSA; and (f) *negative predictive power* refers to the probability that students identified as successful on the CBM measure will be similarly identified as successful on the PSSA.

As evident, there is a trade-off between sensitivity and specificity. One can only be increased at the expense of the other. To set cut scores that maximize each of these measures, a series of receiver operating characteristics (ROC) curves were developed that modeled the

diagnostic accuracy of the CBM and PSSA over a range of cut scores (Swets, 1996). This analysis identifies the point at which sensitivity is maintained with little decrease in specificity. The results of this analysis are seen in Table 5, which found that scores in the winter of 125 wcpm for District 1 and 126 wcpm for District 2 obtained the highest levels of both sensitivity and specificity. In math, scores in computation for District 1 of 8 digits correct in spring and 14 dcpm in District 2 in winter attained the highest levels of diagnostic accuracy. Likewise, a score of 37 in the spring for District 1 in concepts/application also showed the highest levels of diagnostic accuracy.

Outcomes With Published Norm-Standardized Tests

Pearson product-moment correlations were calculated between the CBM reading, math computations, and math concepts/applications scores and the standard scores of subtests of each of the standardized achievement tests administered to students in each district (see Tables 6, 7, and 8). As seen in Table 6 for District 1 at Grade 4, the MAT-8 showed moderate to strong correlations with CBM reading scores across all subtests. On the Total Reading, Sounds and Print, Vocabulary, and Comprehension subtests, correlations ranged from .519 to .724 across assessment periods, with most correlations .633 or better. Somewhat lower correlations were found between CBM reading scores and the Open-Ended Reading subtest, a measure that uses a rubric to judge student performance. These same findings were evident for fifth-grade students in District 1, where correlations between the SDRT and CBM reading scores were .524, .518, and .551 across the fall, winter, and spring CBM assessment periods. Correlations for CBM reading for District 2 with the SAT-9 are shown in Table 7 and show consistent moderate to strong relationships between CBM and all subtests of the norm-referenced standardized test.

Correlations between outcomes for District 2 and the SAT-9 mathematics subtests are shown in Table 8. With the exception of the fourth grade fall data, all correlations were statistically significant and ranged from .45 to .72. Errors in scoring or administration rendered the fall fourth grade data as not useable.

Discussion

CBM With PSSA

Overall, the results of this study showed that CBM reading measures had moderate to strong relationships with the state high-stakes assessment measure, with correlations close to and approaching .70. The outcomes were consistent for both third and fifth grades in both districts in which the normative data were collected. In particular, the CBM measures obtained during the winter or spring assessment period were the strongest contributors to the outcomes of the PSSA. These findings are very consistent with previous research reported across multiple states with CBM reading measures (Powell-Smith, 2004) and add Pennsylvania to the growing number of states where CBM reading measures have shown to be good to excellent predictors of student outcomes on state achievement tests. Considering that each state assessment measure is typically built to evaluate student progress toward competency on state curriculum standards, and that these standards vary considerably from state to state, CBM is indeed a very powerful measurement tool that appears to transcend the differences in state assessments.

Table 5
Diagnostic Accuracy Across Both Districts

	DISTRICT 1											
	Benchmark Period				Benchmark Period				Benchmark Period			
	Content Area/Measurement: Reading/wcpm				Content Area/Measurement: Math Comprehension/dcpm				Content Area/Measurement: Math Concepts/Total Points			
	Fall	Winter	Spring		Fall	Winter	Spring		Fall	Winter	Spring	
Cut score	109	125	142		6	7	8		30	34	37	
Sensitivity (true positive rate)	.76	.75	.70		.53	.64	.71		.53	.66	.76	
Specificity (true negative rate)	.75	.72	.70		.50	.67	.79		.50	.66	.74	
False positive rate	.25	.28	.30		.50	.33	.21		.50	.33	.26	
False negative rate	.24	.25	.30		.47	.36	.29		.47	.33	.24	
Positive predictive power	.86	.84	.83		.68	.79	.88		.68	.80	.85	
Negative predictive power	.60	.58	.53		.35	.48	.58		.35	.50	.61	
Hit rate	.76	.74	.70		.52	.65	.74		.52	.66	.75	
Kappa	.48	.53	.37		.03	.28	.46		.03	.30	.47	
Phi	.48	.44	.38		.03	.29	.47		.03	.31	.48	
	DISTRICT 2											
	Benchmark Period				Benchmark Period				Benchmark Period			
	Fall	Winter	Spring		Fall	Winter	Spring		Fall	Winter	Spring	
Cut score	120	126	135		9	14	19		9	14	19	
Sensitivity (true positive rate)	.69	.86	.80		.62	.77	.77		.62	.77	.77	
Specificity (true negative rate)	.67	.83	.78		.59	.76	.76		.59	.76	.76	
False positive rate	.33	.17	.22		.41	.24	.24		.41	.24	.24	
False negative rate	.31	.14	.20		.38	.23	.23		.38	.23	.23	
Positive predictive power	.85	.94	.91		.82	.91	.91		.82	.91	.91	
Negative predictive power	.43	.68	.58		.33	.52	.52		.33	.52	.52	
Hit rate	.68	.86	.80		.61	.77	.77		.61	.77	.77	
Kappa	.30	.65	.53		.16	.46	.46		.16	.46	.46	
Phi	.32	.66	.54		.18	.48	.48		.18	.48	.48	

Note: wcpm = words correct per minute; dcpm = digits correct per minute.

Table 6
Pearson Product Moment Correlations for CBM Reading and MAT-8 Scores
for Fourth Graders in District 1

	Fall	Winter	Spring
Total Reading	.724	.708	.701
Sounds and Print	.543	.529	.519
Vocabulary	.667	.633	.638
Comprehension	.669	.665	.653
Reading—Open Ended	.457	.412	.421

Note: CBM = curriculum-based measurement; MAT-8 = Metropolitan Achievement Test—Eighth Edition. All correlations significant at $p < .001$.

Table 7
Pearson Product Moment Correlations for CBM Reading and SAT-9 Scores
for Second and Fourth Graders in District 2 Across Assessment Periods

		Fall	Winter	Spring
Total Reading	2nd grade	.711	.737	.740
	4th grade	— ^a	.637	.623
Word Study	2nd grade only	.438	.466	.464
	2nd grade	.704	.702	.733
Vocabulary	4th grade	— ^a	.599	.574
	2nd grade	.718	.744	.695
Comprehension	4th grade	— ^a	.592	.577

Note: CBM = curriculum-based measurement; SAT-9 = Stanford Achievement Test—Ninth Edition. All correlations were significant at $p < .001$.

a. No data were available to conduct these analyses.

Table 8
Pearson Product Moment Correlations for CBM Math and SAT-9 Scores
for Second and Fourth Graders in District 2 Across Assessment Periods

		Fall	Winter	Spring
Total Math	2nd grade	.543	.646	.534
	4th grade	.058	.634	.688
Problem Solving	2nd grade	.508	.623	.527
	4th grade	.029	.529	.575
Procedures	2nd grade	.499	.564	.456
	4th grade	.084	.687	.727

Note: CBM = curriculum-based measurement; SAT-9 = Stanford Achievement Test—Ninth Edition. All correlations are significant at $p < .001$ except fall for fourth grade, which were nonsignificant.

A particularly important contribution of this study was the inclusion of outcomes related to CBM measures of mathematics computation and concepts/applications and state assessments. Results of this study found good support for the use of CBM computation as a moderate predictor of outcomes on state assessments. The findings for math computation were consistent across winter and spring assessment periods, grades, as well as across two school districts. Although the correlations were not as strong as for reading, the correlations were consistently in the .50 range, and all were statistically significant. The relationship of CBM math concepts/applications to the PSSA was also moderate to strong in the winter and spring assessment periods and grades, with most correlations exceeding .60. These data show relationships that were similar to those found in reading. Considering that there is little or no research reported on the relationships of math CBM to state assessment outcomes, the results of this study are quite encouraging in adding the use of the CBM math measures as predictors of student outcomes on state achievement measures.

To further determine how well CBM measures predict outcomes on the PSSA, an analysis of diagnostic accuracy was conducted for each of the measures. For reading, math computation, and math concepts/application CBM, the findings were consistent. Using ROC curves to establish cut points, all measures showed positive predictive power (i.e., correct prediction that a student who was below criterion on CBM was below criterion on PSSA) of around 80% to 93%, and negative predictive power (i.e., correct prediction that a student who was above criterion on CBM was above criterion on PSSA) of around 48% to 68%. Overall correct classification rates for reading and math were between 66% and 85%, a level that suggests that the CBM metrics are good measures for uses of screening. Both specificity and sensitivity across districts were found to be above .7 or .8 in reading and .6 or .7 in math, levels considered acceptable for screening purposes.

CBM With Standardized Tests

Additional analysis of the data examined the relationships between CBM measures and norm-referenced, standardized achievement tests given as part of the districtwide annual testing program. Results of these analyses were very consistent with the decades-long findings that CBM measures show moderate to strong relationships to outcomes on standardized tests (e.g., Shinn, 1989). In reading, correlations between CBM oral reading fluency and both the MAT-8 and SAT-9 were found to be in the .70s. Of particular note were the correlations between the reading comprehension subtests of the standardized tests and the CBM measure, which ranged from .653 to .744. Given that CBM reading is continually criticized for its failure to assess reading comprehension, the results of this study continue to reinforce the concurrent validity of the procedure of having students read a passage aloud for 1 minute as an excellent indicator of overall reading comprehension skills. Additionally, outcomes for math computation and concepts/applications also showed very strong relationships to the SAT-9 in one district and are consistent with findings reported by Helwig et al. (2002), one of the very few studies to report outcomes for CBM mathematics and state high-stakes assessment measures.

Implications for Practice and Research

A potential implication of this study is the possible use that CBM may have in serving as an effective screening device that predicts outcomes on statewide assessment measures. The

findings of this study showed that a 1-minute sample of reading, or a 3- to 5-minute sample of completing math computation problems obtained during the month of January or February, had very strong predictive power to identify those students who were likely to meet or exceed the proficient criterion on the state assessment administered 2 to 3 months after the CBM assessment data were collected. Although the predictive value of the CBM measure did contain a number of false positive and false negative decisions, the measure could offer districts an inexpensive and efficient mechanism to potentially identify a large group of students who were at risk for not being successful on the statewide assessment measure. Such identification could lead a district to designing an intensive, short-term remediation program focused on teaching students the skills and competencies needed to be successful on the statewide assessment. Such remediation efforts would be important for students as well as districts, given the high-stakes nature of these statewide assessment measures. Further studies that examine the longitudinal predictive power of CBM are certainly needed, although some research has already begun to establish that outcomes of 1-minute reading samples obtained at the end of first grade are predictive of performance on statewide high-stakes tests at Grade 3 (e.g., Hintze & Silbergitt, 2005; Keller & Shapiro, 2005).

It is also important to note the efficiency of these CBM measures as compared to norm-referenced achievement tests. Given the expense and time required to administer norm-referenced achievement tests, CBM offers a potentially inexpensive way for districts to do large-scale screening. In addition, because norm-referenced achievement tests were never designed to be responsive to short-term instructional interventions, the use of CBM measures can serve the added purpose of allowing teachers to monitor student performance across time if an intervention plan targeted at students who are at risk for failing statewide assessment measures is implemented. The results of this study suggest that the CBM measures in reading and math can serve as powerful predictors, capturing a high percentage of those students who are likely to be unsuccessful on the statewide assessment.

Another implication of this study is the relationship between CBM measures and acquisition of state standards. The moderate to strong correlations between these measures suggests that the acquisition of state standards through the instructional process is reflected in the CBM measure. Although the CBM measure only appears to assess one skill (i.e., oral fluency) and does not have content validity to the state assessment measure, the CBM measure certainly has strong concurrent validity. Given the efficiency and ease of the measure, CBM measures can play an important role for schools in the screening process.

Limitations

This study does have a number of limitations that require cautious interpretation of findings. First, the study was conducted in only two school districts in Pennsylvania. Although these two districts represented very different demographics (i.e., one was a mixed socioeconomic status [SES], urban/suburban district with a relatively high level of limited English-proficient [LEP] learners and the other was a suburban, high-SES district with a relatively low level of LEP learners), there is certainly the possibility that the outcomes of these districts would not be fully representative of other districts within the state of Pennsylvania. Additional replication of the findings of these districts across other districts in Pennsylvania is certainly needed.

Second, only students who had full data sets (i.e., fall, winter, spring CBM, PSSA, standardized assessments) were included in the analysis. As such, some attrition from the fall sample of CBM data to the final data set was inevitable. Such attrition can potentially result in a final-analysis sample that was somewhat different from the original sample and presents a limitation to interpreting the data.

Third, as noted in the Results section, there were possible problems with the way CBM normative data were collected or scored at one point in time (i.e., fourth grade, District 2). Correlations for math computation in the fall period appeared substantially lower than correlations for other time periods for the same academic skill. Anecdotal, post hoc discussions with the data collectors did not suggest any differences in the way the data were collected or scored, but certainly the data outcomes from those assessment points were not consistent with other data points. Replication efforts across other districts are needed to determine if the effects found among the fourth-grade data collection were idiosyncratic to this study or represent a broader issue about the performance of fourth-grade students at the fall normative assessment point.

Finally, this study offers findings related to performance in mathematics computation and concepts/applications. As this is one of the first studies to report these types of data, it is unknown whether the findings here are similar to what would be found in other types of predictive validity studies. Further studies examining CBM mathematics performance and statewide assessment measures are certainly needed.

Conclusions

Despite the limitations of the study, the results found were consistent with studies in other states. The remarkably strong results across states, especially in reading, suggests that CBM reading is a measure that relates very well to the outcomes of state assessments. As such, student performance on the 1-minute reading passage is clearly related to the large number of skills embedded in the standardized test. In addition, this study is one of the first to show that CBM assessments of math computation and concepts/applications have equally strong predictability to state assessment outcomes.

Overall, this study continues to demonstrate the value of GOMs, and CBM in particular, for evaluating student performance in reading and mathematics. Future studies need to replicate these efforts across districts within Pennsylvania and other states. Given the state-to-state differences in the makeup of assessment measures, such validation is needed. In addition, the long-term predictability of CBM has yet to be determined. This study found that CBM outcomes collected in the winter of the year in which the statewide assessment was given can offer moderate to strong relationships to data collected in the spring of the same year. Can such predictions be made more long term, such as across years? Such data would be important to understanding the degree to which one can identify students at risk for failure on statewide assessment measures long before the student actually takes the test.

This study adds to the empirical research base supporting the use of CBM as a predictor of statewide assessment in reading. The study also contributes to the knowledge base in mathematics, as it offers one of the first pieces of evidence that CBM assessment in math has equally strong predictive power. Future studies and replications of these findings are needed.

References

- Barger, J. (2003). *Comparing the DIBELS oral reading fluency indicator and the North Carolina end of grade reading assessment* (Technical Report). Asheville: North Carolina Teacher Academy.
- Braden, J. P. (2002). Best practices for school psychologists in educational accountability: High stakes testing and educational reform. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (Vol. 1, pp. 301-320). Bethesda, MD: National Association of School Psychologists.
- Buck, J., & Torgeson, J. (2003). *The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test* (Technical Report 1). Tallahassee: Florida Center for Reading Research.
- Castillo, J. M., Torgeson, J. K., Powell-Smith, K. A., & Al Otaiba, S. (2003). *Relationships of five reading fluency measures to reading comprehension in first through third grade*. Manuscript in preparation.
- Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7*, 303-323.
- Deno, S. D. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.
- Deno, S. D. (2003). Developments in curriculum-based measurement. *Journal of Special Education, 37*, 184-192.
- Deno, S. L., Espin, C. A., & Fuchs, L. S. (2002). Evaluation strategies for preventing and remediating basic skill deficits. In M. R. Shinn, H. M. Walker, & G. Stoner (Eds.), *Interventions for academic and behavior problems II: Preventative and remedial approaches* (pp. 213-241). Bethesda, MD: National Association of School Psychologists.
- Edformation. (2005). *AIMSweb progress monitoring and improvement system*. Available from <http://www.aimsweb.com/>
- Erikson, R., Ysseldyke, J., Thurlow, M., & Elliott, J. (1998). Inclusive assessments and accountability systems: Tools of the trade in educational reform. *Teaching Exceptional Children, 31*(2), 4-9.
- Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11*, 513-516.
- Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*, 188-193.
- Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children, 57*, 488-500.
- Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28*(4), 659-671.
- 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*(1), 27-48.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.
- Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1998). *Monitoring Basic Skills Progress—Basic math computation* (2nd ed., Blackline Masters). Austin, TX: Pro-Ed.
- Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1999). *Monitoring Basic Skills Progress—Basic math concepts and applications* (2nd ed., Blackline Masters). Austin, TX: Pro-Ed.
- Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 61-88). New York: Guilford.
- Good, R. H., Simmons, D. C., & Kame'enui, 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*(3), 257-288.
- Harcourt Brace Educational Measurement. (1996). *Stanford Achievement Test* (9th ed.). San Antonio, TX: Harcourt Assessment.
- Harcourt Brace Educational Measurement. (2000). *Metropolitan Achievement Test* (8th ed.). San Antonio, TX: Harcourt Assessment.
- Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *Journal of Special Education, 36*(2), 102-112.

- Hintze, J. M., & Silbergitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review, 34*(3), 372-386.
- Impara, J. C., & Plake, B. S. (Eds.). (1998). *The thirteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
- Karlsen, B., & Gardner, E. F. (1995). *Stanford Diagnostic Reading Test* (4th ed.). San Antonio, TX: Harcourt Assessment.
- Keller, M. A., & Shapiro, E. S. (2005, April). *General outcome measures and performance on standardized tests: An examination of long-term predictive validity*. Paper presented at the annual meeting of the National Association of School Psychologists, Atlanta, GA.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.
- Marston, D. (1989). A curriculum-based approach to assessing academic performance: What it is and why do it. In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford.
- McGlinchey, M. T., & Hixon, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*, 193-203.
- Mead, R., Smith, R. M., & Swanlund, A. (2003, December). *Technical analysis: 2003 Pennsylvania System of School Assessment, mathematics and reading*. Retrieved June 1, 2005, from the Pennsylvania Department of Education Web site: <http://www.pde.state.pa.us>
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, H.R. 1, 115 Stat. 1425 (2002, January 8).
- Pennsylvania Department of Education. (2002). *Handbook for report interpretation: 2002 PSSA mathematics and reading assessment for grades 5, 8, and 11*. Harrisburg, PA: Author.
- Pennsylvania Department of Education. (2003, February). *Technical analysis: Pennsylvania system of school assessment, 2002 reading and mathematics PSSA*. Harrisburg, PA: Author.
- Powell-Smith, K. A. (2004, February). *Individual differences in FCAT performance: A national context for our results*. Paper presented at the annual meeting of the Pacific Coast Research Conference, Coronado, CA.
- Shaw, R., & Shaw, D. (2002). *DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado State Assessment Program (CSAP)* (Technical Report). Eugene: University of Oregon Press.
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.
- Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford.
- Shinn, M. R., & Bamonto, S. (1998). Advanced applications of curriculum-based measurement: "Big ideas" and avoiding confusion. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 1-31). New York: Guilford.
- Shinn, M. R., Shinn, M. M., Hamilton, C., & Clarke, B. (2002). Using curriculum-based measurement in general education classrooms to promote reading success. In M. R. Shinn, H. M. Walker, & G. Stoner (Eds.), *Interventions for academic and behavior problems II: Preventive and remedial approaches* (pp. 113-142). Bethesda, MD: National Association of School Psychologists.
- Sibley, D., Biwer, D., & Hesch, A. (2001). [CBM and its relationship to state assessment in Illinois]. Unpublished data. Arlington Heights, IL: Arlington Heights School District 25.
- Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.
- 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*(3), 407-419.
- Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics*. Hillsdale, NJ: Lawrence Erlbaum.
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26.
- Thurlow, M. L., & Thompson, S. J. (1999). District and state standards and assessments: Building an inclusive accountability system. *Journal of Special Education Leadership, 12*(2), 3-10.