

What is Measured in Mathematics Tests? Construct Validity of Curriculum-Based Mathematics Measures

Robin Schul Thurber
Puyallup School District

Mark R. Shinn
University of Oregon

Keith Smolkowski
Oregon Research Institute

Abstract. Mathematics assessment is often characterized in the literature as being composed of two broad components: Computation and Applications. Many assessment tools are available to evaluate student skill in these areas of mathematics. However, not all math tests can be used in formative evaluation to inform instruction and improve student achievement. Mathematics curriculum-based measurement (M-CBM) is one tool that has been developed for formative evaluation in mathematics. However, there is considerably less technical adequacy information on M-CBM than CBM reading. Of particular interest is the construct that M-CBM measures, computation or general mathematics achievement. This study utilized confirmatory factor analysis procedures to determine what constructs M-CBM actually measures in the context of a range of other mathematics measures. Other issues examined in this study included math assessment in general and the role of reading in math assessment. Participants were 207 fourth-grade students who were tested with math computation, math applications, and reading tests. Three theoretical models of mathematics were tested. Results indicated that a two-factor model of mathematics where Computation and Applications were distinct although related constructs, M-CBM was a measure of Computation, and reading skill was highly correlated with both math factors best fit the data. Secondary findings included the important role that reading skills play in general mathematics assessment.

A large number of students receive special education services for academic achievement problems. According to the *19th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act* (IDEA; U.S. Department of Education, 1997), over 5 million students were served under

IDEA during the 1995-96 school year. More than half of those students receiving special education services (2,597,231) were identified with achievement deficits, and therefore, served under the learning disability category. Even more students are at-risk for developing academic problems (National Center for Edu-

Authors' Notes. This article is based on the doctoral dissertation of the first author and was supported in part by grant No. 84.029D60057 Leadership Training in Curriculum-Based Measurement and Its Use in a Problem-Solving Model sponsored by the US Department of Education, Office of Special Education Programs. The views expressed within this paper are not necessarily those of the USDE. Address all correspondence and questions about this manuscript to Mark R. Shinn, Ph.D., College of Education, University of Oregon, Eugene, OR 97403. E-mail: mshinn@oregon.uoregon.edu

cational Statistics, 1996). When combined with the large number of students served in special education nationally, almost one student in five receives some type of remedial education to reduce academic achievement deficits (Shinn & McConnell, 1994).

Traditionally, reading deficits have received the most attention in the education literature. For example, in a review of the literature on reading disabilities, Hallahan, Kaufman, and Lloyd (1985) asserted that reading is essential for academic functioning in nearly all subjects. However, these researchers also suggested that the common assumption that disabilities in *mathematics* are not as prevalent as those in reading and writing is misleading, if not completely false. Numerous evaluations reveal significant deficits in mathematics performance for students in the United States (National Assessment of Educational Progress [NAEP], 1992; Reese, Miller, Mazzeo, & Dossey, 1997). For example, more than 80% of eighth-grade students could not solve modestly difficult problems (e.g., compute with decimals, fractions, and percents; recognize geometric figures; solve simple equations) correctly from their eighth-grade math textbook (NAEP, 1992). Anrig and LaPointe (1989) found that only 16% of eighth-grade students in the U.S. mastered the content of the typical eighth-grade mathematics text. NAEP results also revealed that only 8% of eighth graders could answer mathematics questions requiring problem-solving skills (NAEP, 1992). Finally, the latest National Assessment of Educational Progress in Mathematics reported that across Grades 4, 8, and 12, 25% or fewer students were estimated to be at the Proficient level or beyond, where students should demonstrate evidence of solid academic performance. Only 2 to 4% of students attained the Advanced level, where students should demonstrate superior performance (Reese et al., 1997).

Formative Evaluation and Mathematics Curriculum-Based Measurement

It has been demonstrated repeatedly (Fuchs & Fuchs, 1986; Fuchs, Fuchs, & Hamlett, 1989; Fuchs, Fuchs, Hamlett, &

Stecker, 1990) that part of effective intervention is formative evaluation. In contrast to summative evaluation that is retrospective (i.e., data are collected *after* completion of instruction), formative evaluation involves the collection of data *during* instruction as a basis for modifying that instruction (Deno & Espin, 1991). By formatively evaluating students' mathematics progress, teachers can assess the effectiveness of their instruction within weeks to determine if their programs are working (Deno, 1986).

Because formative evaluation requires repeated, frequent measurement by classroom teachers, the measurement procedures must be technically sound, quick and easy to administer and interpret, and yield useful information about student performance in basic skills (Deno, 1985; Shinn, 1989). Curriculum-based measurement (CBM) possesses these features. CBM is a well-established technology for measuring student proficiency in reading (Deno, 1985; Shinn, 1989, 1998). Less is known about the technical adequacy of math curriculum-based measurement (M-CBM) where students write answers to standardized computation tasks drawn from the annual general curriculum on tests that vary from 2-5 minutes. Of the few studies that have been conducted, M-CBM has demonstrated high interrater agreement (.97), high 1-week test-retest reliability (.87), and moderate alternate form reliability (.66; Tindal, Marston, & Deno, 1983).

In M-CBM validity studies, the emphasis has been on concurrent validity. A relation with commercial norm-referenced math tests provides only modest support for validity. Few reported correlations exceed .60 and the median correlation is .43 with the Problem-Solving subtest and .54 with the Math Operations subtest of the Metropolitan Achievement Tests (MAT; Marston, 1989; Putnam, 1989).

Two hypotheses have been offered to explain these lower than expected correlations (Marston, 1989). First, the limited content validity of the criterion commercial mathematics tests (Freeman et al., 1983) may make them inadequate criterion measures. Second, these criterion math tests could be measuring more than just mathematics skills because many of

the items rely on silent reading of the instructions and problems. Thus, reading skills may influence performance on the mathematics test (Skiba, Magnusson, Marston, & Erickson, 1986 as cited in Marston, 1989).

Although some criterion-related validity evidence has been provided for Math-CBM, in and of itself, this type of validity is considered necessary but not sufficient to establish a measure's technical adequacy (Messick, 1990). *Construct validity* is the most important type of validity evidence. Some construct validity evidence (e.g., discriminant validity) for CBM mathematics measures has been demonstrated. For example, Shinn and Marston (1985) found that Math-CBM probes differentiated students in general education, Title 1, and programs for mild disabilities at Grades 5 and 6. Students with mild disabilities also were distinguished from general education students in Grade 4. However, a preferable way to evaluate construct validity is to examine the factor structure of *multiple* measures (Good & Jefferson, 1998). By using confirmatory factor analysis, the correlation of each measured variable with the constructs (e.g., math competence) shared by the measures can be estimated. This approach has been used to support the validity of CBM reading measures (Shinn, Good, Knutson, Tilly, & Collins, 1992) but has not been used to examine the construct validity of CBM mathematics measures.

What Construct Does M-CBM Measure?

Although different names may be used, two broad constructs of mathematics performance are typical in the education literature: computation or operations, and applications or "problem solving." Computation involves working math problems where students must know the concepts, strategies, and facts (Howell, Fox, & Morehead, 1993; Silbert, Carnine, & Stein, 1990). Applications are the use and understanding of math concepts to solve problems (e.g., applied word problems, measurements, temperature, volume (Salvia & Ysseldyke, 1991). This problem solving is the functional combination of computation and application knowledge (Howell et al., 1993).

M-CBM was designed to serve as a measure of *general* math achievement, not specifically as a measure of only computation or applications. This theory is predicated on the hypothesized high relation between computation and application. The purpose of this study was to examine the relation of M-CBM to the constructs of general mathematics achievement, computation, and application from a theoretical perspective using confirmatory factor analysis. Three models were tested:

1. A unitary model where Computation and Applications comprise a general math competence construct that M-CBM measures accurately;
2. A two-factor model where Computation and Applications are distinct constructs and M-CBM is a measure of Computation; and
3. A two-factor model where Computation and Applications are distinct and M-CBM is a measure of Applications.

The role of reading in mathematics also was examined in each of the proposed math models.

Method

Participants

Participants were 207 fourth graders from general education classrooms in four elementary schools located in a mid-sized Northwestern public school district. General demographic information on the schools involved in the study is presented in Table 1. Gender was distributed about equally with 46% female and 54% male students. Most participants (74%) received all their instruction in general education; 18% received Title I services, and 8% received special education services in a resource room for part of the day.

Participants were obtained in accordance with University of Oregon protection of human subjects practices. Permission to conduct the study was obtained at the district administrative level and four elementary school principals were provided a written description of the study. Fourth-grade teachers then were contacted and they sent a description of the study and passive consent letters to 213 parents. No parents refused. Complete data were obtained on 207

Table 1
School Demographic Information

School	Total Enrollment	Ethnicity	Socioeconomic Status (SES) ^a	State-Wide Testing ^b
A	564	82% White/non-Hispanic 2% Black/African American 12% Hispanic/Latino	64%	44% Math 48% Reading
B	488	88% White/non-Hispanic 1% Black/African American 8% Hispanic/Latino	46%	56% Math 59% Reading
C	256	94% White/non-Hispanic <1% Black/African American <1% Hispanic/Latino	15%	56% Math 73% Reading
D	96	98% White/non-Hispanic <1% Black/African American	13%	41% Math 50% Reading

^aSES based on percent of students receiving free/reduced lunch. ^bState-wide testing percentiles indicate percent of fifth-grade students meeting or exceeding state standards.

participants as six students were absent from at least one of the three testing sessions.

Measures

Participants were administered the 12 mathematics measures shown in Table 2 selected to provide multiple measures of the hypothesized constructs of computation, applications, or general mathematics competence.

Mathematics CBM probes (M-CBM). M-CBM consisted of three fourth-grade-level math probes sampled from the annual curriculum of typical mathematics texts. The problems required a range of computation skills, from basic addition, subtraction, multiplication, and division facts to more complex use of algorithms and strategies (e.g., 362×25). Each basic skill problem area comprised approximately 8% of the test items, and approximately 36% of the total. The more complex computational problem constituted about 64% of the test items. Participants were given 5 minutes on each probe to complete as many problems as possible. Problems were scored by counting the number of correct digits (CD) in the process of obtaining the answer and the answer itself. As discussed earlier, some evidence

suggests that M-CBM is reliable with respect to interrater scoring and alternate forms, with modest, but limited evidence of validity.

Basic math fact probes. Students were administered two math fact probes containing a combination of addition, subtraction, multiplication, and division facts. Addition and subtraction facts formed approximately 25% of the test items and were distributed equally. Basic addition facts included mathematics problems with whole numbers under 10 (e.g., $0 + 0$ to $10 + 10$). Basic subtraction facts included problems in which the subtrahend (i.e., subtracted number) and the difference (i.e., answer) are single-digit numbers (e.g., $1 - 0$ to $10 - 9$). Approximately 75% of the test items were basic multiplication and division facts that again, were divided equally. Basic multiplication facts consisted of mathematics problems with single-digit factors (e.g., 0×0 to 9×9). Basic division facts included problems in which the divisor and quotient are single-digit numbers (e.g., $0 \div 0$ to $81 \div 9$) (Silbert et al., 1990). Students were given 2 minutes to complete as many problems as possible. Probes were scored by counting the number of problems correct.

Table 2
Measures

Name of Test	Math Computation	Math Applications	Reading
Curriculum-Based Measurement (M-CBM)	3 mixed-operation probes		
Basic Math Facts	2 fact worksheets		
Stanford Diagnostic Mathematics Test (SDMT)	Computation subtest	Applications subtest	
California Achievement Tests (CAT)	Math Computation subtest	Math Concepts and Applications subtest	
National Assessment of Educational Progress (NAEP)		Applications items	
Reading Maze Test			3 Maze tests

Stanford Diagnostic Mathematics Test. Students also were given the Stanford Diagnostic Mathematics Test (SDMT, Beatty, Gardner, Madden, & Karlsen, 1985) Computation and Applications subtests of the Green Level test, intended to be used with students in Grades 4 and 5, and with low-achieving students in Grade 6. The Computation subtest assessed knowledge of the facts and algorithms of addition, subtraction, multiplication, and division, and methods for solving simple and compound number sentences. The Applications subtest assessed skill in applying basic math facts and principles. Items ranged in difficulty from requiring students to solve simple story problems and to select models for solving one-step problems to those that require solving multiple-step and measurement problems. Internal consistency reliability estimates for subtests by grade typically exceeded .90. Criterion-related validity evidence is reported in the test manual with correlations between SDMT subtests and total test score and between the Stanford Achievement Test subtest and total score ranging from .64 to .89.

California Achievement Tests (CAT). Students were administered the Mathematics Computation and Mathematics Concepts and Applications subtests from the CAT (CTB/McGraw-Hill, 1992). The Mathematics Computation subtest assessed skill in solving addition, subtraction, multiplication, and division problems involving whole numbers, fractions, mixed numbers, decimals, and algebraic expressions. The Mathematics Concepts and Applications subtest assessed skill in understanding and applying a variety of mathematical concepts involving numeration, number sentences, problem solving, and measurement. Reliability evidence is restricted to internal consistency with all but two coefficients exceeding .80. Validity evidence for the CAT is limited to a demonstration that the percentage of students mastering objectives increases with age and that the CAT is correlated with the Test of Cognitive Skills.

The National Assessment of Educational Progress (NAEP). Students also completed a set of fourth-grade NAEP items that measured mathematics applications. The NAEP mathematics assessment was designed

to report the progress of students nationally at Grades 4, 8, and 12 based on a framework influenced by the National Council of Teachers of Mathematics (NCTM; 1989). The NAEP purports to examine mathematical abilities (conceptual understanding, procedural knowledge, and problem solving) and mathematical power (reasoning, connections, and communication) but no technical adequacy information is available.

CBM Maze test. Students completed three CBM Maze reading tests (Fuchs & Fuchs, 1992; Shinn, 2002) as a measure of general reading achievement. Each test consisted of a reading passage of approximately 250 words with every 7th word deleted. Students selected responses from three choices with two distracters and only one word that correctly completed the sentence. Participants were given 5 minutes to complete each test. Research has demonstrated the technical adequacy of the CBM Maze test as a valid measure of reading with correlations with commercial reading tests ranging from .60 to .90, with an average correlation of .74 (Fuchs & Fuchs, 1992).

Procedures

Training of data collectors. Seven graduate students from a school psychology program at a major Pacific Northwestern university were trained as data collectors in four 1-hour training sessions. During the first session, they were trained to give M-CBM measures, basic skill probes, and CBM Maze tests with scripted directions after modeling by the first author; this was followed by a second session during which the scoring procedures were taught. The third and fourth sessions covered administration and scoring of the SDMT subtests and the NAEP math test.

Data collection and scoring. All tests were administered to groups of students in their own classrooms. The CAT was given by students' classroom teachers as part of the district's evaluation program. The NAEP was administered by the researchers according to the primary investigator's modifications of the original NAEP directions. The original NAEP

directions included extensive explanations for administering the NAEP as part of a comprehensive nationwide assessment of student learning. Because most of these directions were not applicable to the purpose of this study, these explanations were dropped.

Testing was completed in three sessions across 3 days in a prescribed order. The first session consisted of M-CBM and Maze tests. The second session consisted of the SDMT Computation subtest and NAEP items. The third session consisted of Basic Math Fact probes and the SDMT Concepts and Applications subtest. Each session lasted approximately 45 minutes.

Interrater Agreement

The seven data collectors and the primary investigator independently scored six of the tests for 10 participants: (a) M-CBM, (b) Basic Facts worksheet, (c) Maze Task, (d) SDMT Computation subtest, (e) SDMT Concepts and Applications subtest, and (f) NAEP items. Interrater agreement coefficients of .83, .90, .94, .94, .85, and .77 for total scores were obtained for M-CBM, Basic Facts, Maze Task, SDMT Computation, SDMT Concepts and Applications, and NAEP, respectively, using the formula: $\text{number of agreements} / (\text{number of agreements} + \text{number of disagreements}) \times 100$.

Results

Means and standard deviations for all measures are reported in Table 3. Overall, for tests with multiple forms, scores appeared similar with respect to means and standard deviations. However, the mean of the third M-CBM probe was lower than the other two M-CBM probes (i.e., 61.0, 65.2, and 49.9).

The correlation matrix for all measures is reported in Table 4. Correlations between parallel forms of the same measure were generally high, suggesting high alternate form reliability. For example, correlations among the three M-CBM correlations ranged from .90 to .92.

When examining the correlations among the math tests for evidence of concurrent validity, nearly all correlations were greater than .50. However, some general patterns are no-

Table 3
Descriptive Statistics for Student Performance on
Measured Variables ($n = 207$)

Variable	Mean	Standard Deviation (SD)
Reading Maze 1	32.7	12.4
Reading Maze 2	37.2	13.6
Reading Maze 3	40.7	14.9
M-CBM 1	61.0	33.0
M-CBM 2	65.2	34.2
M-CBM 3	49.9	31.0
Basic Facts 1	31.2	13.9
Basic Facts 2	30.1	13.8
SDMT Comp	14.2	4.4
SDMT App	19.8	6.5
NAEP	17.4	9.2
CAT Comp	28.1	8.9
CAT Comp	29.4	9.9

Note. All scores are reported in raw score units. Maze scores reflect number of correct words circled; CBM scores reflect number of correct digits; Basic Fact scores reflect number of correct basic facts. M-CBM = Curriculum-Based Measurement Math; SDMT = Stanford Diagnostic Mathematics Test; NAEP = National Assessment of Education Progress; CAT = California Achievement Tests.

ticeable. First, the measures typically conceived as measuring Computation correlated more highly with other computation measures. They also correlated lower with measures traditionally conceived as Applications. The reverse pattern was apparent with the measures conceptualized as Applications. With respect to the specific measures of interest in this study, M-CBM, these tests correlated most highly with the computation tasks tested in the Basic Facts 1 and Basic Facts 2 probes, with a median correlation of .82. M-CBM correlated less well with applications measures such as SDMT Applications, CAT Applications, and the NAEP with median correlations of .44.

Because confirmatory factor analysis assumes multivariate normality, a preliminary analysis of the sample data was conducted to test for skewness and, more importantly, kur-

tosis. This test indicated positive skewness greater than 1.0 for the three M-CBM computation probes with coefficients of 1.32, 1.32, and 1.41, respectively. Positive kurtosis greater than 1.0 was found for Maze 1, and the three M-CBM probes with coefficients of 1.19, 1.94, 2.45, and 2.25.

Model Testing

The three models of interest were tested using the *Mplus* statistical analysis package (Muthén & Muthén, 2001). Because of the moderate deviations from normality for some of the measures discussed earlier, models estimated with maximum likelihood estimation were compared to the same models estimated with the robust estimators provided in *Mplus* (Muthén & Muthén, 2001). The fit statistics and parameter estimates differed only slightly

Table 4
Correlations Among Measured Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Maze1													
2. Maze2	.88												
3. Maze3	.87	.88											
4. M-CBM1	.66	.59	.68										
5. M-CBM2	.68	.60	.67	.92									
6. M-CBM3	.65	.57	.65	.90	.91								
7. Basic Facts1	.67	.63	.67	.82	.83	.82							
8. Basic Facts2	.62	.59	.61	.80	.81	.82	.92						
9. SDMT Comp	.57	.54	.57	.58	.59	.54	.67	.61					
10. SDMT App	.58	.55	.58	.42	.42	.36	.51	.47	.66				
11. NAEP	.61	.60	.63	.44	.44	.38	.52	.45	.60	.81			
12. CAT Comp	.64	.63	.66	.62	.63	.59	.66	.62	.82	.69	.66		
13. CAT App	.61	.60	.63	.50	.51	.44	.55	.50	.68	.80	.80	.78	

Note. M-CBM = Curriculum-Based Measurement Computation; SDMT = Stanford Diagnostic Mathematics Test; SDMT1 = Computation subtest; SDMT = Applications subtest; NAEP = National Assessment of Education Progress; CAT = California Achievement Tests; CAT Comp = Computation subtest; CAT App = Concepts/Applications subtest.

(i.e., less than .01) within the levels of precision typically reported. Because differences were small and showed no substantive differences, the models reported were estimated with standard maximum likelihood.

Based on the recommendations noted by Bollen and Long (1993), multiple measures were used to evaluate fit of the hypothesized models. The Tucker-Lewis Index (TLI), also called the Nonnormal Fit Index and ρ_2 in Bollen, 1989) favors more parsimonious models, is sample-size independent, and has been recommended by Marsh (1995). The Comparative Fit Index (CFI), a truncated version of the Relative Noncentrality Index, offers an index of fit when parsimony is less important (Marsh, 1995). Both indices are normed so that they conform to a standard metric, ranging from 0 to 1, and for adequate fit, indices exceeding .95 were expected. These measures, along with the traditional χ^2 and the Root Mean Square Error of Approximation (RMSEA), provide a fairly complete descrip-

tion of model fit. The RMSEA characterizes acceptable fit when it is below .05.

Each model tested in this study specified estimated and constrained parameters. The estimated parameters include factor loadings, residual (error) variances, and correlations that were expected to be nonzero. Constrained parameters include those factor loadings and correlations that were specified to be 0.0 for each model. All parameters are reported as standardized coefficients. For example, factor loadings are reported as correlations between measured variables with the latent factor(s) in the model.

Because of hypothesized confounds of method variance observed in previous model testing with Reading-CBM (Shinn et al. 1992), each model tested included a specific method variance factor as a number of the mathematics measures (including M-CBM and the Maze Reading tests) were very short, timed tests. This Timed Test factor was defined so that it was uncorrelated with content factors, and thus it

should measure variance associated *only* with test methods. In addition, two pairs of measures not associated with the Timed Test factor were related beyond that captured by content factors. The two alternate forms of basic facts were allowed to correlate as were the two CAT measures. Like the timed tests, these pairs of measures were highly similar to each other in the testing methods, yet different from other measures. It is important to account for extraneous sources of variance, such as that associated with test methods, in the models to allow the substantive factors for mathematics and reading to capture the true variance associated with those skills.

Results of model testing are represented in Figures 1 through 3. Latent constructs are displayed in ellipses, and measured variables are portrayed as squares. Directional arrows from the factors to the measured variables represent factor loadings. Curved double-headed arrows between factors or residuals indicate correlations.

Unitary model of mathematics assessment. In general, most of the indices indicated that this model was not a good fit to the data. The chi-square goodness-of-fit was significant, $\chi^2(54) = 190.05, p < .01$, indicating poor fit, and the CFI and TLI were .96 and .94, respectively, indicating a marginal fit to the data. The RMSEA was .110, well above the preferred .05 level.

Two-factor model of mathematics assessment. The unitary model of mathematics, presented in Figure 1 was hierarchically related to the two-factor model displayed in Figure 2 where Computation and Applications are distinct constructs and M-CBM is a measure of Computation. The unitary model is nested within the two-factor model in that the more restrictive unitary model was obtained by applying two constraints on the more general two-factor model (i.e., making Computation and Applications correlate 1.00 and reading correlate equally with both Computation and Applications).

Results indicate that this two-factor model provided an acceptable fit to the data. The chi-square goodness-of-fit was barely sig-

nificant, $\chi^2(52) = 77.23, p = .013$. This indicates that the departure from the observed covariance matrix from the covariance matrix specified by the model may be due to chance sampling variability. The CFI and TLI, however, were both .99, indicating excellent fit to the data. In addition, the RMSEA was .048, another indication of acceptable fit.

In this model, Computation and Applications were distinct, but highly related, .83. Factor loadings on Computation ranged from .60 for CBM 3 to .93 for CAT Computation. Factor loadings on applications ranged from .89 for SDMT Applications to .90 for NAEP and CAT Applications. Reading performance was highly correlated with both Computation ($r = .76$) and Applications ($r = .77$).

Because the unitary model was nested within the two-factor model, a chi-square difference test was calculated to determine which model was an improved fit to the data. The difference in chi-square was significant, $\chi^2(2) = 112.82, p < .01$, indicating that this two-factor model was a substantially better fit to the data.

M-CBM as a measure of Applications. Because there was theoretical and practical interest in using M-CBM to make generalizations to math applications performance, a two-factor model in which Computation and Applications were distinct math constructs and M-CBM was a measure of Applications was tested. As shown in Figure 3, reading again was highly correlated with both Computation (.69) and Applications (.76). A particularly strong correlation was found between Computation and Applications (.88). When factors are so highly correlated, it is difficult to determine that they are measuring distinct constructs. Therefore, it appears that when M-CBM is loaded on the Applications factor, the Computation and Applications constructs cannot be distinguished.

Like all the models, the chi-square goodness-of-fit statistic for this two-factor model was significant, $\chi^2(52) = 133.66, p < .01$. The CFI and TLI were .98 and .96, indicating acceptable fit to the data, but the RMSEA was .087, above the .05 criterion. Based on the chi-square and other measures of goodness-of-fit,

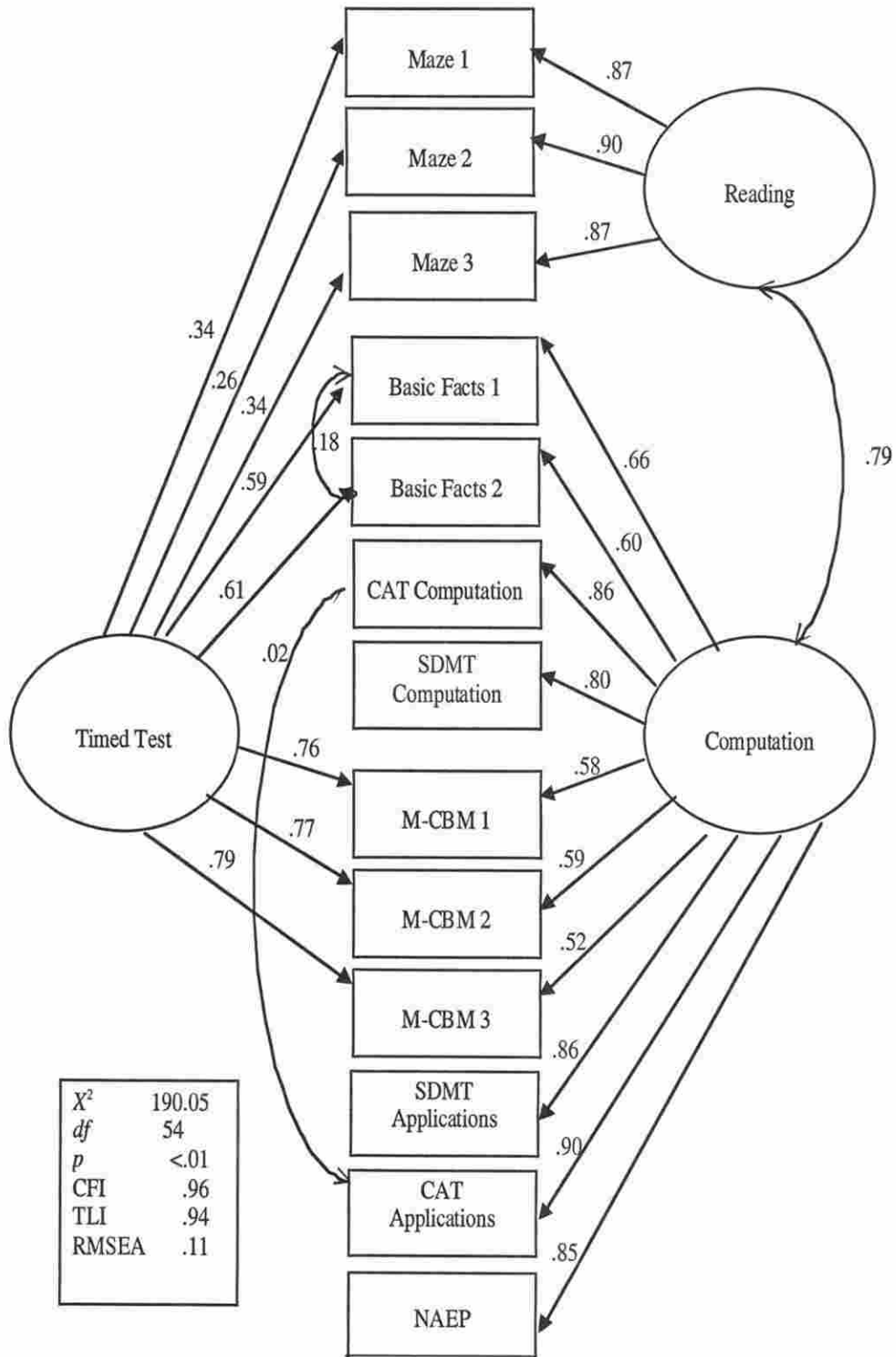


Figure 1. A single-factor model of mathematics assessment inclusive of timed method variance and reading skill.

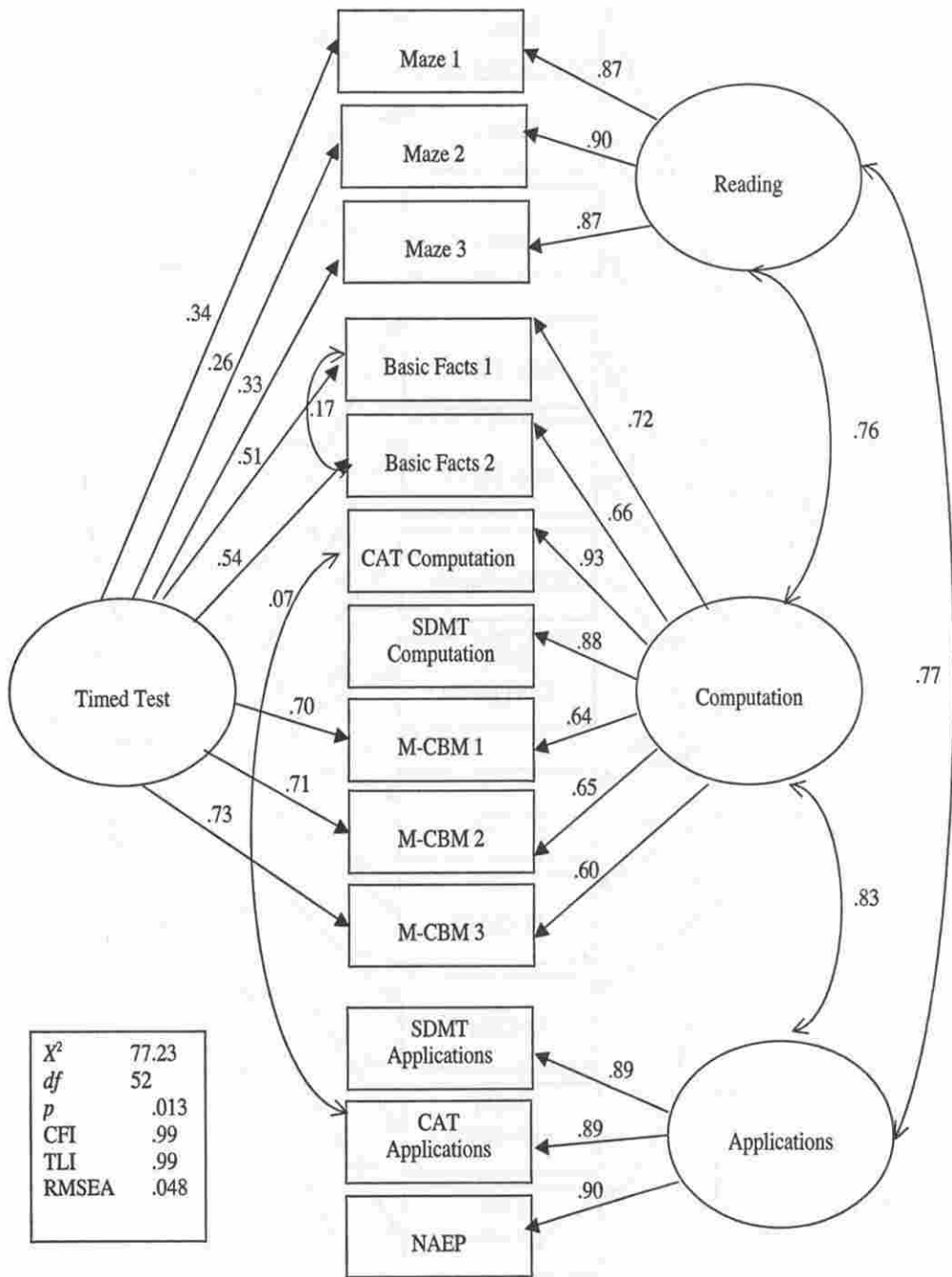


Figure 2. A two-factor model of mathematics assessment inclusive of timed method variance and reading skill with M-CBM as a measure of computation.

this two-factor model with M-CBM measuring Applications was *not* a better fit to the data than the two-factor model with M-CBM measuring Computation.

Discussion

M-CBM was developed to address the need for ongoing progress monitoring in mathematics computation. Despite a plethora of research conducted to validate the use of CBM reading, and to a lesser degree CBM written expression and spelling, the technical adequacy of the M-CBM has not been investigated thoroughly (Marston, 1989; Putnam, 1989; Shinn, 1989). Important questions remain unanswered regarding which aspect(s) of mathematics, if any, are measured validly by M-CBM. The current study used confirmatory factor analysis procedures in an attempt to answer this question as well as the potential influence of reading in mathematics assessment.

In this study, evidence of high alternate form reliability for M-CBM was observed with a median correlation of .91 among the three forms. However, lower than expected interscorer agreement coefficients of .83 also were observed. Given that examiners used the same scoring key, this low interscorer agreement suggests that counting all the correct digits in the answer is more complex than it appears or that more scorer training is necessary.

Some convergent and divergent validity data for M-CBM also were generated. In examining the correlation matrix in Table 4, M-CBM correlated highly with other measures of basic facts computation (median $r = .82$) and more modestly with commercial measures of math Computation (median $r = .61$). Performance on M-CBM also was less related to tests conceptualized as measuring math applications (median $r = .42$).

Results of model testing indicate that the most defensible model was the one displayed in Figure 2. This model specified a *two-factor* model of mathematics assessment where Computation and Applications were distinct, although *highly related* constructs ($r = .83$). Although the χ^2 statistic was significant for this model, all other reported fit measures indicate a *good* fit to the data.

In this two-factor model, the median factor loading of M-CBM on the Computation construct was .64 providing moderate evidence of its validity as a measure of mathematics Computation. Reading skill, as measured by CBM maze also correlated highly with both the Computation and Application constructs (.76 and .77, respectively). In addition, using the χ^2 difference test and other goodness-of-fit indices, a significant improvement in fit was *not* obtained with other models tested. Therefore, this model cannot be rejected as the most plausible explanation of the data.

Implications

Although mathematics theory is not as well documented as reading theory, two broad factors of math emerge from the literature, Computation and Applications. In practice, these two factors typically are viewed as independent. For instance, this multidimensional theory of math is evident in the scope and sequence of traditional mathematics curricula. Typical math textbooks usually contain "(1) problem sets in which only computations are performed and (2) word problems that require selection and application of the correct algorithm and computation" (Salvia & Ysseldyke, 1991, p. 554). Math tests are constructed along these theoretical lines. It may be important to understand *the degree of dependence* among these constructs, however. It appears that skills in one area are necessary for success in the other.

What also has been ignored typically in mathematics assessment is the role of reading. This role not been well researched, although speculation abounds, especially with respect to math applications. For example, Salvia and Ysseldyke (1991) caution, "Although reading level is popularly believed to affect the difficulty of word problems, its effect is less clearly established" (p. 554). Similarly, Skiba et al. (1986; as reported in Marston, 1989) and Marston, (1982; as reported in Good & Jefferson, 1998) hypothesized that commercial math tests could be measuring *more* than just mathematics skill. Skiba and colleagues (as reported in Marston, 1989) suggested that validity coefficients among math measures were

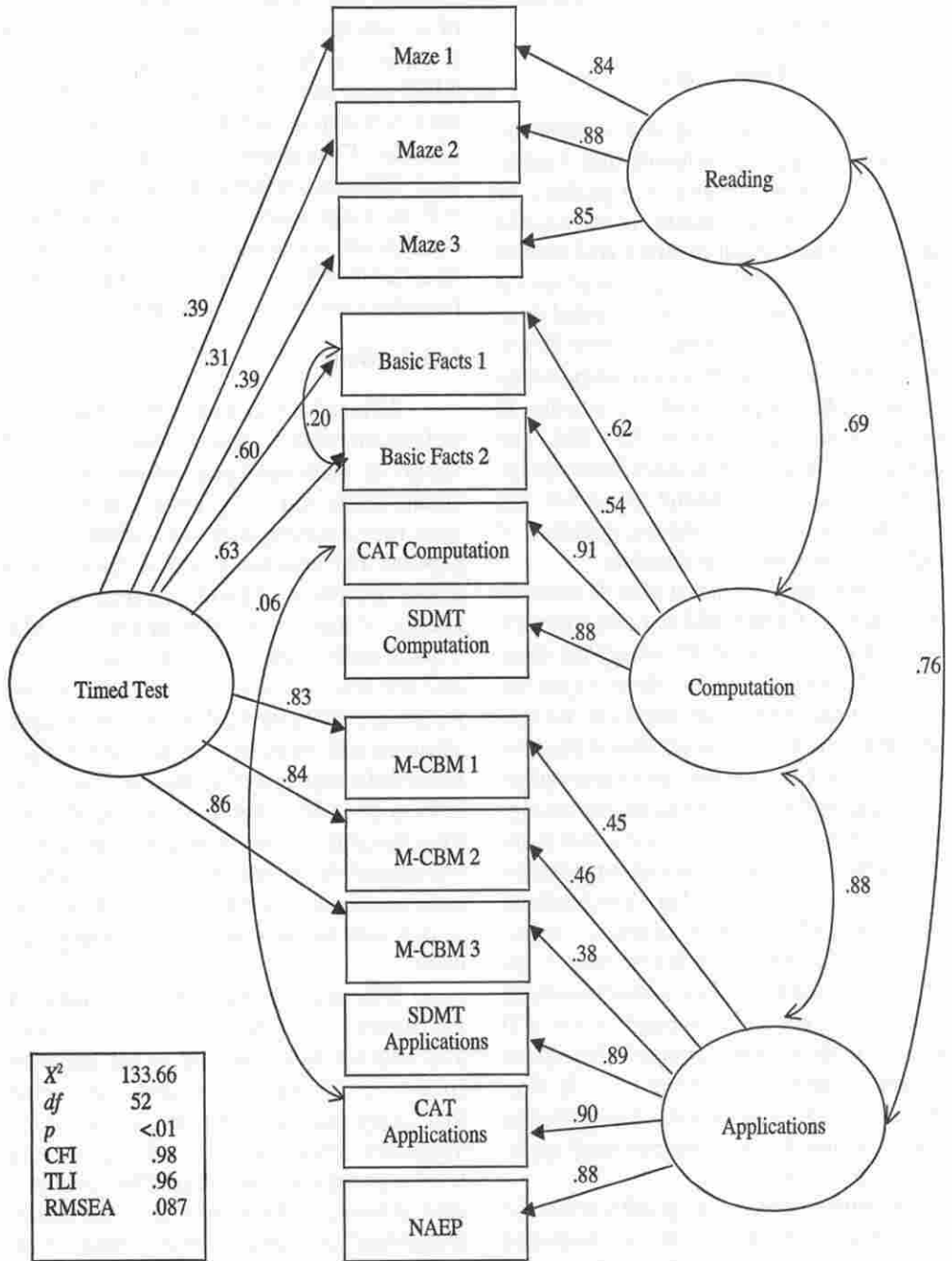


Figure 3. A two-factor model of mathematics assessment inclusive of timed method variance and reading skill with M-CBM as a measure of application.

improved when reading competence was included in a prediction equation. However, the study in which these data were reported was unobtainable and thus could not be verified.

In Figure 2, correlations between the Reading and Computation (.76) and Applications constructs (.77) were high, sharing about one-half of their variance. This suggests that students who performed well in reading also tended to perform well in mathematics. Conversely, students who were not proficient in reading did not perform well on the math measures. Therefore, results indicate that reading may be a necessary and important component in overall math competence and should not be overlooked in drawing conclusions about mathematics skills.

Limitations of the Study

As with all research studies, this study possesses several limitations that affect the interpretations and generalizations of the reported results. Foremost among these limitations is lack of external validity. An effort was made to collect data across a diverse sample of students in terms of gender, educational program, and ethnicity. Regardless, students in this sample represented primarily White/non-Hispanic, general education students from schools from one school district in one Northwestern state. Finally, as only fourth-grade students served as participants, potential developmental differences in mathematics models among different grades cannot be estimated. This study should be replicated with students who differ on these demographic features. Another concern, as is usually the case in confirmatory factor analysis, is sample size. Currently, there is no consensus on the optimal sample size for a confirmatory factor analysis study. Fassinger (1987) reported estimates ranging from 100 for a small study to 30 participants per measured variable for larger studies. Another guideline for the minimum number of participants is 5 to 10 times the number of observed variables (Bryant & Yarnold, 1997). Finally, Marsh, Balla, and McDonald (1988) found that the effect of sample size was still significant for sample sizes as large 400 to 1,600 individuals. Using the aforementioned guidelines, the

sample size of the current study, 207 individuals, is considered adequate by some standards but too small by other standards.

The final major threat to the study's results is the failure to obtain a nonsignificant χ^2 statistic for any of the three models. In conventional interpretation (Bollen, 1989; Bryant & Yarnold, 1997; Fassinger, 1987) this failure suggests that each model did not reproduce the observed data accurately. As mentioned previously, however, there is considerable debate as to whether the χ^2 statistic should be used as the primary or sole indicator of model fit due to a number of limitations (e.g., too restrictive, sensitivity to sample size and multivariate normality). It is possible that a nonsignificant χ^2 could have been obtained by adding additional parameters (e.g., correlated errors, additional loadings), thereby improving the fit. However, it was believed that the mildly significant χ^2 that was obtained by testing a theory-driven model was better than a nonsignificant χ^2 that may have been attained by capitalizing on chance.

Summary

Theories are not proven true but are confirmed or disconfirmed by converging evidence (Stanovich, 1992). This is an important caution in interpreting theoretical research such as the current study. Confirmatory factor analysis, by nature, is designed to *reject* theories, rather than prove them true. Instead, confirmatory factor analysis procedures determine whether the sample data *confirm* the hypothesized models, thus lending support to a proposed theory (Long, 1983). Although this study can be useful in making statements about this sample under these conditions, additional research is necessary to allow for the convergence of information.

The intent of the present study is to *begin* to explore the technical properties of M-CBM and other variables in mathematics assessment. Three outcomes strongly suggest the need for additional research to improve the technical properties of M-CBM. First, as was noted earlier, the interscorer agreement (.83) was too low, given that the examiners were given a scoring key. However, examiners were asked to count the number of digits in the an-

swer and in the process of obtaining the answer. The latter task may either require more examiner inferences that need to be trained and practiced with consistency, better scoring keys, or a simplified scoring scheme.

Second, there was evidence of both positive skew and kurtosis. The former suggests that the grade-level computation tasks may have been too difficult for the students. It is worth noting that M-CBM was the only non-“broadband” mathematics measure. All the others sampled a range of across-grade types of problems. With respect to the kurtosis, the reported value suggests a restricted range.

Third, the relation of M-CBM to the Computation factor, although defensible as evidence of validity, was lower than expected. Whether this correlation reflects the cumulative effects of low interscorer agreement, skewness, and kurtosis, or accurately reflects the relation of the type of measure to the construct should be investigated.

References

- Anrig, G. R., & LaPointe, A. E. (1989). What we know about what students don't know. *Educational Leadership*, 47(3), 4-9.
- Beatty, L. S., Gardner, E. G., Madden, R., & Karlsen, B. (1985). The Stanford Diagnostic Mathematics Test (3rd ed.). San Antonio, TX: The Psychological Corporation.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.
- Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models*. Newbury Park, CA: Sage Publications, Inc.
- Bryant, F. B., & Yarnold, P. R. (1997). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99-136). Washington, DC: American Psychological Association.
- CTB/McGraw-Hill. (1992). *California Achievement Tests, Fifth Edition*. Monterey, CA: CTB Macmillan/McGraw-Hill.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, 52, 219-232.
- Deno, S. L. (1986). Formative evaluation of individual programs: A new role for school psychologists. *School Psychology Review*, 15, 358-374.
- Deno, S. L., & Espin, C. A. (1991). Evaluation strategies for preventing and remediating basic skill deficits. In G. Stoner, M. R. Shinn, & H. M. Walker (Eds.), *Interventions for achievement and behavior problems* (pp. 79-97). Silver Spring, MD: National Association of School Psychologists.
- Fassinger, R. E. (1987). Use of structural equation modeling in counseling psychology research. *Journal of Counseling Psychology*, 34, 425-436.
- Freeman, D. J., Juhn, T. M., Porter, A. C., Floden, R. E., Schmidt, W. H., & Schwille, J. R. (1983). Do textbooks and tests define a national curriculum in elementary school mathematics? *Elementary School Journal*, 83, 501-513.
- Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53, 199-208.
- Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review*, 21, 45-58.
- Fuchs, L. S., Fuchs, D., & Hamlett, C. L. (1989). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education*, 10, 43-52.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review*, 19, 6-22.
- Good, R. H., & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M. R. Shinn (Ed.), *Advances in curriculum-based measurement* (pp. 61-88). New York: Guilford Press.
- Hallahan, D. P., Kaufman, J. M., & Lloyd, J. W. (1985). *Introduction to learning disabilities*. Englewood Cliffs, NJ: Prentice-Hall.
- Howell, K. W., Fox, S. L., & Morehead, M. K. (1993). *Curriculum-based evaluation: Teaching and decision making*. Pacific Grove, CA: Brooks/Cole.
- Long, J. S. (1983). *Confirmatory factor analysis*. Newbury Park, CA: Sage Publications.
- Marsh, H. W. (1995). χ^2 and χ^2 I2 fit indices for structural equation models: A brief note of clarification. *Structural Equation Modeling*, 2, 246.
- Marsh, H.W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 391-410.
- Marston, D. B. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford Press.
- Messick, S. (1990). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012-1027.
- Muthén, L. K., & Muthén, B. O. (2001). *Mplus user's guide*. Los Angeles: Muthén & Muthén.
- National Assessment of Educational Progress. (1992). *NAEP 1992 mathematics report card for the nation and the states* (Report No. 23-ST02). Washington, DC: National Center for Educational Statistics.
- National Center for Educational Statistics. (1996). *The pocket guide to the condition of education: 1996*. Washington, DC: U.S. Department of Education.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.

- Putnam, D. (1989). *The criterion-related validity of CBM measures of math*. Unpublished master's thesis, University of Oregon, Eugene.
- Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *NAEP 1996 mathematics report card for the nation and the states*. Washington, DC: National Center for Education Statistics.
- Salvia, J., & Ysseldyke, J. E. (1991). *Assessment* (5th ed.). Boston: Houghton Mifflin.
- Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children*. New York: Guilford.
- Shinn, M. R. (Ed.). (1998). *Advanced applications of curriculum-based measurement*. New York: Guilford.
- Shinn, M. R. (2002). *Use of curriculum-based measurement maze in general outcome measurement*. Eden Prairie, MN: Edformation, Inc.
- Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, 21, 459-479.
- Shinn, M. R., & Marston, D. (1985). Differentiating mildly handicapped, low-achieving, and regular education students: A curriculum-based approach. *Remedial and Special Education*, 6(2), 31-38.
- Shinn, M. R., & McConnell, S. M. (1994). Improving general education instruction: Relevance to school psychologists. *School Psychology Review*, 23, 351-371.
- Silbert, J., Carnine, D., & Stein, M. (1990). *Direct instruction mathematics* (2nd ed.). Columbus, OH: Merrill.
- Skiba, R., Magnusson, D., Marston, D., & Erickson, K. (1986). *The assessment of mathematics performance in special education: Achievement tests, proficiency tests, or formative evaluation?* Minneapolis: Special Services, Minneapolis Public Schools.
- Stanovich, K. E. (1992). *How to think straight about psychology*. New York: Harper Collins.
- Tindal, G., Marston, D., & Deno, S. (1983). *The reliability of direct and repeated measurement* (Research Report No. 109). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.
- U.S. Department of Education. (1997). *Nineteenth annual report to Congress on the implementation of the Individuals with Disabilities Education Act*. Washington, DC: U.S. Department of Education.

Robin Schul Thurber received her Ph.D. in School Psychology from the University of Oregon in 1999 and is a school psychologist in Puyallup, Washington. Areas of interest include instructional design and consultation, curriculum-based measurement (CBM), and violence prevention/social skills instruction.

Mark R. Shinn received his Ph.D. in Educational Psychology (School Psychology) from the University of Minnesota in 1981 and is a Professor in the Special Education area at the University of Oregon. His primary research and teaching interests are curriculum-based measurement and its use in a problem-solving model and other needs-based service delivery systems.

Keith Smolkowski currently works as a research analyst at Oregon Research Institute. He received his Master's degree in Decision Sciences in 1995 from the University of Oregon and is pursuing a Ph.D. in Special Education there. His interests include research methods and statistics, effective literacy instruction, and positive behavior support.

