

Technical Adequacy of the Maze Task for Curriculum-Based Measurement of Reading Growth

Jongho Shin, Stanley L. Deno, and Christine Espin, *University of Minnesota*

The purpose of the present study was to examine the technical adequacy of curriculum-based measurement (CBM) for assessing student growth over time. Participants were 43 second graders whose reading performance was measured monthly over 1 school year with the maze task. Technical characteristics of the CBM maze task were examined in terms of reliability, sensitivity, and validity for assessing student growth. Results showed that the maze task had good alternate-form reliability, with a mean coefficient of .81 and 1- to 3-month intervals between testing. The maze task also sensitively reflected improvement of student performance over a school year and revealed interindividual differences in growth rates. Finally, growth rates estimated on repeated maze scores were positively related to later reading performance on a standardized reading test; in addition, although a significant difference was not found, general education students appeared to develop reading proficiency faster than remedial education students. Results support the use of the maze task as a reliable, sensitive, and valid data collection procedure for assessing reading growth.

Curriculum-based measurement (CBM) holds promise for assessing student growth over time because of its distinctive characteristics. These characteristics include (a) provision of multiple data points over short time periods, (b) sensitivity to small changes in student performance, and (c) absolute measures of student performance (L. S. Deno, 1985; Marston, Deno, & Tindal, 1983; Marston & Magnusson, 1985). In addition to these characteristics, however, instruments designed to assess growth must be sensitive to academic skill development, and the growth rates estimated on repeated performance measures must be logically related to criterion measures. The purpose of this study was to examine these psychometric characteristics of CBM as a multiwave progress monitoring system.

How to measure students' academic growth over time has been a critical issue in education. Historically, efforts have been made to describe student growth by using the difference score between measures obtained at two points in time (Cronbach & Furby, 1970; D. J. Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1994). Indexing student growth with a difference score, however, has been criticized on the basis of such psychometric problems as unreliability, inaccurate relations to initial status and correlates, and incompatibility

(Bereiter, 1963; Cronbach & Furby, 1970; Lord, 1956, 1963; O'Connor, 1972; Willet, 1989). Fortunately, the development of advanced statistical methods like hierarchical linear models (HLM; Bryk & Raudenbush, 1987, 1992) enables researchers to overcome the technical difficulties of difference scores by allowing multiwave data points to be used for assessing student growth.

To measure academic growth over time, however, there must also be testing instruments that provide multiple data points representing student performance over short time periods (e.g., 1 year). Commonly used instruments such as published standardized achievement tests are not appropriate for repeatedly measuring student performance within a school year. In addition, standard scores derived from these commercial tests are intended to reveal differences between individuals at one time rather than within an individual over time. The use of standard scores also artificially limits increases of individual differences over time, which hinders appropriate representation of academic growth over time (D. J. Francis et al., 1994; Labouvie, 1982; Lord, 1963; Willet, 1989).

In contrast, CBM enables multiple data points to be obtained to scale growth over short time periods in a logistically efficient way (L. S. Deno, 1985, 1989). In addition, CBM is sen-

sitive in detecting small changes in student performance (Marston et al., 1983; Marston & Magnusson, 1985), and CBM measures represent absolute performance levels rather than relative standings among students. These characteristics of CBM make it distinct from other achievement tests in that CBM also meets the basic technical requirements of testing instruments for assessing growth (i.e., interval or ratio scale, the same performance level represented by the same score obtained on different occasions, and no change in the underlying construct being assessed; D. J. Francis et al., 1994; Shin, Deno, Espin, & McConnell, 1999; Willet & Sayer, 1994).

Research has been conducted to examine the technical adequacy of CBM as a progress monitoring system; however, most research has been based on static measures obtained at a single time point, not on growth rates estimated on repeated measures of student performance. For example, studies on the criterion-referenced validity of CBM have been conducted primarily by investigating the relation between CBM and criterion measures obtained on a single occasion (L. S. Deno, 1985; S. L. Deno, Mirkin, & Chiang, 1982; Espin & Foegen, 1996; Good & Jefferson, 1998; Marston, 1989; Marston & Magnusson, 1985; Parker, Hasbrouck, & Tindal, 1992). Although the early findings on CBM support its use as an indicator of general student performance at a given point in time, they do not demonstrate the technical adequacy of the CBM repeated measurement data as a measure of growth assessment.

Several more recent studies have been conducted to examine technical features of the multiple measures of student performance obtained through CBM. For example, L. S. Fuchs and Fuchs (1992) investigated the ratio of the average growth rate estimated by different CBM reading tasks to their standard error of estimation. L. S. Fuchs, Fuchs, Hamlett, Waltz, and Germann (1993) examined the developmental patterns of students' academic skills using linear and quadratic growth terms, the distribution of growth rates estimated on multiwave performance measures, and group differences in growth rates between grades. Although the results of these studies support the use of CBM for assessing academic skill development over time, they do not directly address several important issues related to the multiwave CBM data.

The present study was designed to extend the recent research on the technical adequacy of repeated CBM for quantifying growth. Three specific psychometric characteristics immediately and practically important for scaling growth were examined in the study: (a) alternate-form reliability, (b) likely sensitivity to actual changes in the academic skills being developed, and (c) criterion validity of the growth rates estimated through repeated CBM.

Alternate-form reliability was examined because CBM involves repeated measures of student performance using alternate forms of the test to gather multiple data points. Neither internal consistency reliability nor test-retest reliability is completely informative in revealing the technical adequacy of CBM for assessing growth because these constructs represent

the degree of response consistency to test items and the consistency of relative standings among students, respectively (Crocker & Algina, 1986). Although some early studies examined alternate-form reliability of CBM (Marston, 1989; Parker, Tindal, & Hasbrouck, 1989), alternate-form reliability in these studies was obtained by administering alternate forms to students on the same occasion. If alternate forms are used for assessing academic growth over time, determining alternate-form reliability across time is also important for establishing the technical adequacy of the measure in terms of equivalency of different test forms and stability of the measure itself (Crocker & Algina, 1986; stability is not synonymous with growth—see our later discussion for details).

Second, we examined the likely sensitivity of CBM data to real growth in the academic construct being measured. Certainly, elementary students increase their reading skills as group across a school year (Bast & Reitsma, 1998; H. Francis, 1992; Shin, 1999). At the same time, however, we can assume that the reading proficiency of individuals within groups develops at differential rates. If CBM is to be used for assessing student growth, it must sensitively reflect both group growth and individual differences in reading skill development.

Third, the criterion validity of growth rates estimated through CBM repeated performance measures was examined. If estimated growth rates are to evaluate program effectiveness and make predictions about student outcomes, they must relate to other important measures (e.g., other achievement measures, group differences between general and special education students). Without criterion validity, growth estimates from repeated measures may be practically useless. This examination of the criterion validity of CBM was conducted to extend the earlier work on the criterion validity of single point or current level estimates of performance produced through CBM.

The characteristics examined in the study might be considered to be immediately practical in their importance for using CBM in progress monitoring or assessing growth (D. J. Francis et al., 1994; Willet & Shayer, 1994). In this regard, we are extending the work of L. S. Fuchs and Fuchs (1992) and L. S. Fuchs et al. (1993) in examining the technical adequacy of CBM for assessing growth.

The technical characteristics of CBM examined in this study focused on the maze task, a measure increasingly used in CBM. The maze task is typically constructed by deleting every seventh word and replacing it with three multiple choice alternatives (one correct word and two clearly incorrect words). When taking the CBM maze task, students are asked to select a correct answer from the three alternatives with a time limit (e.g., 3 minutes). The maze task was used in the present study because it offers practical advantages for assessing student growth in reading. Unlike the typical oral reading measure used in CBM, with the maze students can be tested in groups. This makes it time efficient. In addition,

computers can be used to administer and score repeated measures. Another practically important advantage of the maze task is that it has greater face validity for practitioners because it appears to measure reading comprehension as well as word recognition (L. S. Fuchs & Fuchs, 1992).

Method

Participants and Setting

This study was conducted as part of a larger study on the use of technology for facilitating inclusion of students with mild disabilities in general classroom settings. Forty-three second graders (25 male and 18 female students) from three classes in a large urban school in the Midwest participated. Participants included 2 Native American, 12 African American, 2 Asian American, 2 Hispanic, and 25 Caucasian students.

Eighteen students received Chapter I services in reading and mathematics; 1 student was labeled learning disabled, and 24 students did not receive any remedial educational services. All participants took the California Achievement Test (CAT; CTB/ McGraw-Hill, 1985) toward the end of second grade, in April 1996. Mean scaled scores on the reading and mathematics subtests for participants receiving remedial educational services were 652 and 648, corresponding to the 47th and 46th percentiles on the national norms, respectively. In contrast, mean scaled scores for students in general education were 704 and 703 on the reading and mathematics subtests, corresponding to the 71st and 79th percentiles.

Procedure

Ten different maze passages were used to assess students' reading performance over a school year. Passages were randomly selected from generic grade-level reading materials. To construct maze reading tests, every seventh word was deleted after the first sentence, and three alternatives were provided. One of the alternatives was a correct choice and the other two

were distracters. Distracters were designed to be easily distinguished from the correct choice (see Figure 1, and L. S. Fuchs & Fuchs, 1992, for details). The number of correct choices in the maze task was scored and used for the data analysis in the study.

Maze measures were collected monthly, from September through June, using different forms of the task. Data were collected using the discourse system, which is a computer-based classroom communication system where students' own minicomputers are connected to a teacher's computer (see Robinson, DePascale, & Roberts, 1989; Shin, Deno, Robinson, & Marston, 2000, for details). The discourse system provides many educationally useful functions, one of which is administering and scoring a programmed test automatically. The maze passages in the present study were programmed into the discourse system. Segments of the passage appeared on the students' minicomputers, and students were given 3 minutes to read the maze passage and select answers. Students were asked to type in the first letter of their selected answer. At the end of 3 minutes, the discourse system recorded students' responses and scored the correct choices. Students had no difficulty using the discourse system because they had been using it daily throughout the school year during the ordinary classroom instruction.

Data Analysis

Four different analyses were conducted to address the research questions. First, to examine alternate-form reliability of the maze task across time, the correlations between monthly maze scores obtained by different forms of the maze task on each testing were computed using the Pearson product-moment correlation method. The correlation between monthly maze scores with 1- to 3-month intervals between testing was of in be estimated most reliably by using at least 3 data points, distributed as evenly as possible across the school year (Willet, 1989). Given the present data set, the minimum number of 3 equally distributed data points would require testing of at least 3-month intervals. Therefore, the analysis of alter-

**My mother always likes to go home. She was born on a nice (farm/ big/ soon) in a valley.
Her father started (home/ the/ sat) farm before she was born. When (red/ she/ told) was a
little girl they lived (to/ fun/ in) a very old log house on (call/ date/ the) farm. One day her
father said (that/ little/ size) they were going to build a (try/ new/ with) house. That made
my mother, her (before/ longer/ sister), and her brothers very excited. (continued)**

FIGURE 1. A sample maze probe.

nate-form reliability in the study focused on 1- to 3-month intervals between testing.

A second research question focused on both the sensitivity of the maze task for revealing improvement in reading proficiency and interindividual differences in growth rates over a school year. To answer this question, hierarchical linear models (Bryk & Raudenbush, 1987, 1992) were employed:

$$Y_{it} = \pi_{0i} + \pi_{1i} (\text{Month}) + r_{it} \text{ (within-individual model)}$$

$$\pi_{1i} = \beta_{10} + u_{1i} \text{ (between-individual model),}$$

where Y_{it} is the monthly maze score for individual i at time t , π_{0i} is the intercept of the growth line indicating the performance level at the beginning of the school year for individual i , π_{1i} is the monthly growth rate for individual i , r_{it} is the random error associated with individual i , β_{10} is the mean growth rate for the whole group of students, and u_{1i} is the random effect related to the group's mean growth rate. The statistical test on the mean growth rate (β_{10}) shows the sensitivity of the maze task to group improvement in reading proficiency, whereas the statistical test on the variance of individual growth rates shows the sensitivity of the maze task to interindividual differences in rates of developing reading proficiency.

Third, the validity of the maze task was examined by looking at the relation between students' growth rates estimated on repeated maze scores and student performance on the reading subtest of the California Achievement Tests (CAT) using HLM. In this analysis, student performance on the CAT was included in the between-individual model as a level-two predictor: $\pi_{1i} = \beta_{10} + \beta_{11} (\text{CAT})_i + u_{1i}$, where β_{10} is the mean growth rate for students who had the CAT reading scores at the mean and β_{11} is the regression coefficient representing the relation between students' growth rates estimated on the maze task and student performance on the CAT.

Finally, a group difference in growth rates between students who did and did not receive remedial educational services was examined. In this analysis, a between-individual model included the variable of service type as a level-two predictor: $\pi_{1i} = \beta_{10} + \beta_{11} (\text{Service Type})_i + u_{1i}$, where β_{10} is the

mean growth rate for general education students and β_{11} is the mean difference in growth rates between general and remedial education students.

Results

Reliability

The correlation between monthly maze scores with 1- to 9-month intervals ranged from .69 to .91, with a mean of .81 (see Table 1). The correlation between maze scores with 1-month intervals between testing (which is displayed immediately under the diagonal) ranged from .75 to .90, with a mean of .83. The correlation for 2-month intervals ranged from .75 to .87, with a mean of .80, and the correlation for 3-month intervals ranged from .69 to .91, with a mean of .80.

Sensitivity

Descriptive statistics for monthly maze scores are displayed in Table 2. As can be seen, the obtained means and standard deviations for performance on the maze task increased over time. HLM was used to test the reliability of this descriptive pattern.

First, to test the sensitivity of the maze task for detecting improvement of reading proficiency, we examined whether the mean growth rate of the group (β_{10}) was statistically different from null growth. Results showed that the mean growth rate was statistically significant ($\beta_{10} = 1.07$, $t = 11.47$, $p < .01$), meaning that the increase in maze scores was reliable. That obtained mean growth rate was an increase of 1.07 correct choices per month on the maze task. To examine the accuracy in estimation of the mean growth rate, we computed the ratio of the standard error of estimate to the mean growth rate (see L. S. Fuchs & Fuchs, 1992). The standard error of estimation of the mean growth rate was .09. The ratio of the standard error to the mean growth rate was .08. Because we might be concerned if the ratio was greater than, say, .50, we believe that the mean growth rate was estimated reliably.

TABLE 1. Correlation Between Monthly Maze Scores for 1- to 9-Month Intervals Measured with Alternate Forms

	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	March	April	May	June
Sept.	1.00									
Oct.	0.85	1.00								
Nov.	0.81	0.78	1.00							
Dec.	0.91	0.80	0.87	1.00						
Jan.	0.88	0.82	0.80	0.90	1.00					
Feb.	0.82	0.79	0.69	0.87	0.86	1.00				
March	0.76	0.69	0.72	0.77	0.82	0.80	1.00			
April	0.85	0.86	0.89	0.86	0.82	0.77	0.75	1.00		
May	0.84	0.77	0.78	0.83	0.86	0.82	0.75	0.80	1.00	
June	0.79	0.74	0.75	0.82	0.87	0.85	0.80	0.76	0.84	1.00

Note. All correlation coefficients were statistically significant at the level of .01 with a sample size of 38.

TABLE 2. Means and Standard Deviations of Monthly Maze Scores

	Sept.	Oct.	Nov.	Dec.	Jan.	Feb.	March	April	May	June
Mean	4.36	5.19	6.47	8.14	9.78	8.34	11.21	12.38	14.28	10.83
SD	3.93	3.89	4.81	5.90	6.31	5.70	7.19	7.80	8.38	6.99
<i>n</i>	42	42	38	42	40	38	38	42	39	40

Second, the sensitivity of the maze measure for revealing interindividual differences in growth rates was tested by computing the statistical significance of the parameter variance for individual growth rates. The estimated parameter variance of .25 was significantly different from zero, meaning that students differed significantly from one another in their individual growth rates ($X^2 = 128.34$, $df = 42$, $p < .01$). In addition, to examine the accuracy of the estimation of the growth parameter variance, the ratio of the standard error of estimate to the estimated variance was computed. The standard error of estimate of the growth parameter variance was .08. The ratio of the standard error to the growth parameter variance was .32, indicating that the growth parameter variance was estimated somewhat reliably.

Validity

Before examining the relationships between growth rates for monthly maze scores and the criterion measures (i.e., reading performance on the CAT and membership in general or remedial education), we examined the reliability of estimated growth rates. The reliability of the estimates serves as an index of how reliably the relations between CBM and the criterion measures will be examined by HLM (Bryk & Raudenbush, 1987, 1992). The reliability of growth rate estimates is defined as the proportion of the observed variance of growth rates attributable to the true parameter variance. The reliability of estimated growth rates was .66 in this study, indicating that 66% of the variance of growth rate estimates could be attributed to the true variance. We concluded from this that the relations between growth rates and criterion measures could be examined reliably.

We then examined the relationship between growth rates estimated from repeated maze scores and the reading scores on the CAT. Our analysis revealed a significant positive relationship between the two measures ($\beta_{11} = .33$, $t = 3.31$, $p < .01$). Students with CAT reading scores 1 *SD* higher than the mean showed higher growth rates by .33 correct choices per month, on average, than students with the CAT reading scores at the mean.

Subsequently, we tested the group difference in growth rates between general and remedial education students using HLM. The obtained mean growth rate for general education students was an increase of 1.20 correct choices per month,

and the mean growth rate for remedial education students was an increase of .91 correct choices per month. The difference in the mean growth rates between the groups (i.e., .29 increases per month), however, was not statistically significant ($t = 1.51$, $p = .14$). This failure to find differences in growth rates between the two groups was in contrast to their scores at the beginning of the school year. The initial levels of reading performance estimated by HLM was 7.16 correct choices for general education students and 2.67 correct choices for remedial education students ($t = 4.27$, $p < .01$). Figures 2 and 3 show individual differences in growth rates and initial performance levels around the mean values for general and remedial education students, respectively.

Discussion

CBM holds promise for researchers interested in studying student growth and the relation between student growth and other variables. It is easy to develop alternate CBM test forms, the tests are simple and efficient to administer and score, and test results are easy to interpret and communicate to others (L. S. Deno, 1985). However, many questions regarding the technical adequacy of CBM measures for the study of growth as a multiwave growth monitoring system have remained unanswered.

This study investigated the technical adequacy of the CBM maze task for describing growth over time based on repeated measures of student performance. First, alternate-form reliability of the maze task was examined with various time intervals between two equivalent tests. Results showed that the average correlation estimates for 1- to 3-month intervals between testing were in the .80s. Alternate-form reliability estimates in this study are compatible with those in early research on alternate-form reliability (Bradley, Ackerson, & Ames, 1978; Marston, 1989; Parker, Tindal, & Hasbrouck, 1989). Results provide evidence that the maze task reliably assesses student growth in reading based on repeated performance measures collected at intervals from 1 to 3 months over an academic year.

The second question we addressed was whether repeated measures using the maze task were sufficiently sensitive to real student growth over time. We can reasonably assume that overall group performance will increase across an academic

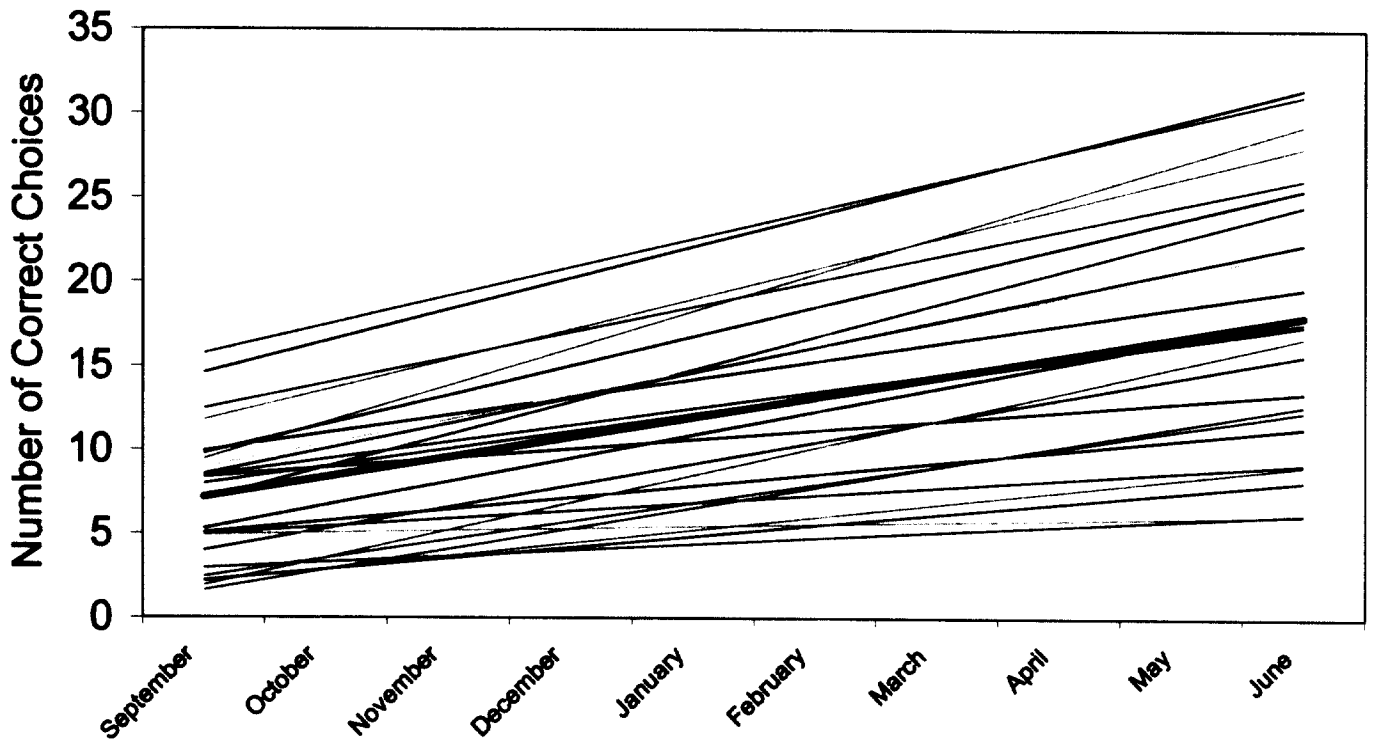


FIGURE 2. Estimated growth lines for individual students in general education. The thick dark line represents the mean growth line for general education students (growth rate of 1.20 increases per month and initial performance level of 7.16).

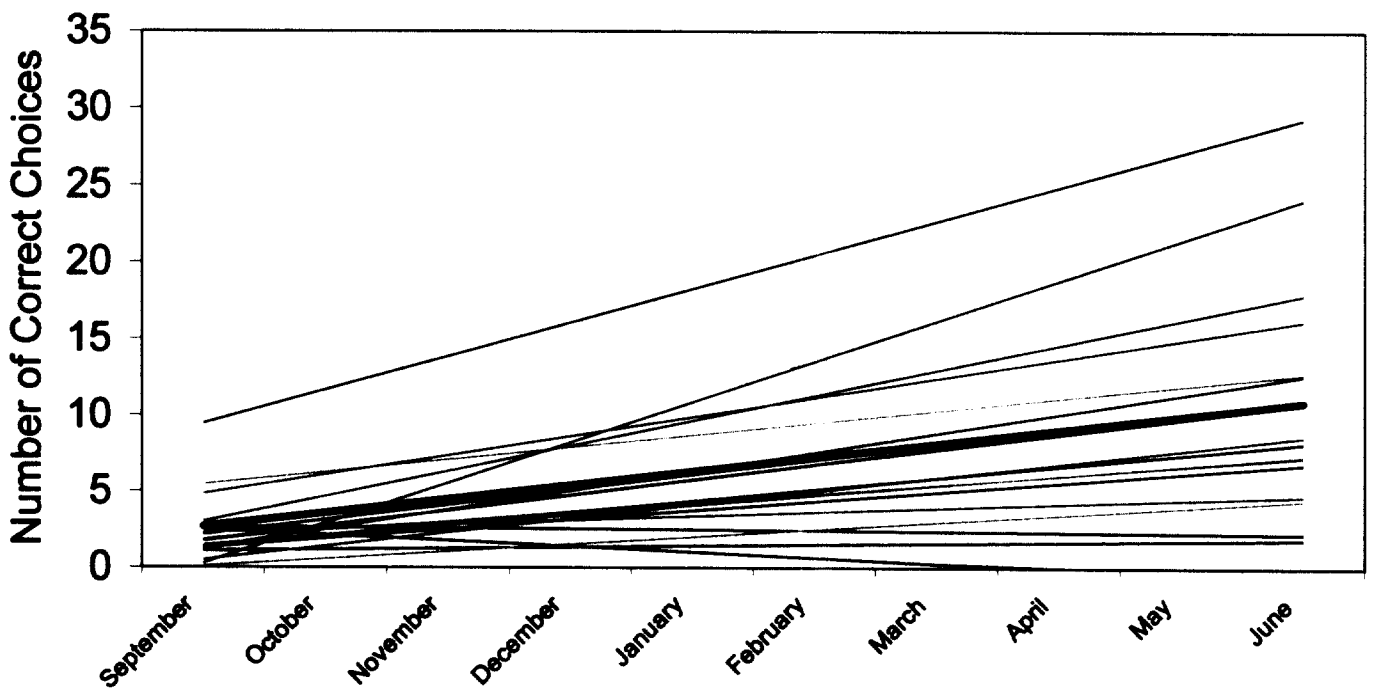


FIGURE 3. Estimated growth lines for individual students in remedial education. The thick dark line represents the mean growth line for remedial education students (growth rate of .91 increases per month and initial performance level of 2.67).

year, especially for students in lower grades who are acquiring basic reading skills in decoding and comprehension (Bast & Reitsma, 1998; H. Francis, 1992; Shin, 1999). At the same time, we assume that all students do not grow at the same rate in learning basic skills (Shin, 1999). If group growth and individual differences are not observed in repeated measurements, the test is not suitable for growth monitoring. Our results revealed that maze performance sensitively reflects significant group improvement as well as interindividual differences in improvement, even for lower grade elementary students (see Jenkins & Jewell, 1993).

The finding of interindividual differences in growth rates also revealed that the variances of monthly maze scores increased over time. Unfortunately, such increases in the variability of student performance over time suggest that the gap between students who are high and low achieving grows larger over time. This is the "Matthew effect" (Stanovich, 1986) so often observed in the early development of reading skills. For educators, this means that programs for students who are slower to learn should be modified as early as possible to boost rates of growth. Future research should be conducted to identify instructional factors that can be used to raise growth rates for these students (e.g., classwide peer tutoring strategies discussed by D. Fuchs, Fuchs, Mathes, & Simmons, 1997; Phillips, Hamlett, Fuchs, & Fuchs, 1993).

Finally, our results support the use of the CBM maze task as predictive of student achievement differences. As in L. S. Fuchs and Fuchs (1992), growth rates estimated from multiwave measures rather than single scores were examined. We found that the growth rates estimated from repeated maze scores were predictive of later student performance on the standardized reading test administered at the end of the school year. We note with interest in this context that the difference in growth rates between general and remedial education students was less significant than the differences in levels of performance at the outset. We wonder whether this result was a statistical artifact or was indicative of successful remedial education.

We need to identify several technical concerns present in this study related to alternate-form reliability, the relation between growth rates and static criterion measures, and relatively lower maze growth rates. With regard to alternate-form reliability, it is possible that a negative relation exists between estimates of growth and alternate-form reliability (i.e., the higher the alternate-form reliability, the smaller the growth, and vice versa). This is an important technical problem when the study of growth is based on only 2 data points. However, in a study using multiple data points, researchers are able to overcome this inverse relationship between growth and reliability of the measure because growth rates are estimated using all available multiwave data points rather than only 2 points (Willet, 1989). It is also possible that the problem does not occur even when 2 data points are used for assessing growth. Because the correlation between two test scores is based on

the consistency of relative rankings and proportional intervals between rankings on the two tests (Howell, 1997), a uniform increase of test scores among individuals (e.g., a gain of five correct choices on the maze task over a week for all students) on two alternate forms of the test results in a perfect alternate-form reliability, even with the existence of a large growth.

An argument could also be made that growth rates need not be positively correlated with static outcome measures because of the regression toward the mean (i.e., a negative relation between initial status and slope). The regression effect becomes a significant problem when the study of growth is based on the two-wave research design (Willet, 1989). In the two-wave research design, students with lower initial status are more likely to have larger gain scores than those with higher initial status, and vice versa. Therefore, it is difficult to expect a positive relation between gain scores and static criterion measures.

This argument, however, does not hold true when the assessment of growth is based on multiple data points, as in the present study (Willet, 1989). The use of multiple data points with an advanced statistical method like HLM enables researchers to estimate growth rates and examine the relationship between initial status and slope more accurately by controlling for the regression effect (Bryk & Raudenbush, 1987, 1992; Willet, 1989). Therefore, it is reasonable to expect a positive relation between growth rates estimated on multiple data points and static criterion measures as validity evidence of the measure for assessing growth. In the present study, the estimated correlation between initial status and growth rates was .97 using HLM. The high positive correlation also indicates the existence of the "Matthew effect" (Stanovich, 1986) in the early development of reading proficiency.

Finally, the low maze growth rates could be viewed as a concern for progress monitoring. The magnitudes of maze growth rates are low (e.g., 1.07 increases per month for second graders in this study) relative to those obtained through reading aloud (e.g., 9.14 increases per month for second graders converted from the weekly growth rate reported by L. S. Fuchs et al., 1993). As a result, the maze task might not appear to be as sensitive in assessing growth over time as reading aloud. L. S. Fuchs and Fuchs (1992), however, found that there was compatibility in growth rate estimates between the maze and reading aloud tasks, and that maze growth rates could be converted to reading aloud growth rates. Consequently, when a large number of students are monitored, the maze task is advantageous over the reading aloud task because of its characteristics of group administration and computer use, as shown in the present study.

With respect to the primary questions of the study, we conclude that performance scores from the maze task are technically adequate and can be used in CBM to assess student growth and monitor progress in reading performance. We believe, though, that further research on student growth with

multiwave maze measures is needed to (a) identify the growth patterns for students receiving general and special education services (Shin, 1999), (b) estimate typical growth rates for these two populations at different grades (L. S. Fuchs et al., 1993; Shin, 1999), and (c) establish functional relations between growth rates and instructional or individual characteristic variables. Using CBM measures including the maze task could allow researchers to study student growth in a more reliable, sensitive, and valid way because of its distinctive technical characteristics and efficiency in developing alternate test forms and administering tests frequently.

We would like to close with a comment related to the differences between general and remedial education students. In our study, the remedial education students performed around the average level in the national norm on the standardized achievement test administered at the end of the school year; yet, their reading performance at the beginning of the school year on the maze task was significantly lower (2.67 correct choices) than that of general education students (7.16 correct choices). We are uncertain whether the year-end performance of these students on the standardized achievement test reflects the effect of extra-educational services provided during the school year, or reflects that the groups do not differ. In any case, the extra-educational services provided for remedial education students very likely confound the examination of differences in growth rates between general and remedial education students in terms of validity evidence for the maze task as a measure of assessing growth.

Ultimately, we believe that educational decisions for individual students are more dependable if they are based on growth measures with multiwave data points than static scores obtained on single occasions, especially for students receiving special education services. We hope that more research conducted to study student growth and CBM can contribute to broadening current research practices in this regard.

REFERENCES

- Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Results from a Dutch longitudinal study. *Developmental Psychology, 34*, 1373-1399.
- Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring* (pp. 3-20). Madison: University of Wisconsin.
- Bradley, J. M., Ackerson, G., & Ames, W. S. (1978). The reliability of the maze procedure. *Journal of Reading Behavior, 10*, 291-296.
- Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*, 147-158.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace.
- Cronbach, L. J., & Furby, L. (1970). How we should measure "change"—Or should we? *Psychological Bulletin, 74*, 68-80.
- CTB/McGraw-Hill. (1985). *California achievement test*. Monterey, CA: Author.
- Deno, L. S. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.
- Deno, L. S. (1989). Curriculum-based measurement and special education services: A fundamental and direct relationship. In M. R. Shinn (Ed.), *Curriculum-based measurement* (pp. 1-17). New York: Guilford.
- Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36-45.
- Espin, C. A., & Foegen, A. (1996). Validity of general outcome measures for predicting secondary students' performance on content-area tasks. *Exceptional Children, 62*, 497-514.
- Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1994). Measurement of change: Assessing behavior over time and within a developmental context. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities* (pp. 29-58). Baltimore: Brookes.
- Francis, H. (1992). Patterns of reading development in the first school. *British Journal of Educational Psychology, 62*, 225-232.
- Fuchs, D., Fuchs, L. S., Mathes, P. G., & Simmons, D. C. (1997). Peer-assisted learning strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34*, 174-206.
- 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., Waltz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*, 27-48.
- 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.
- Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury.
- Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59*, 421-432.
- Labouvie, E. W. (1982). The concept of change and regression toward the mean. *Psychological Bulletin, 92*, 251-257.
- Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 421-435.
- Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring* (pp. 21-38). Madison: University of Wisconsin.
- Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why it is. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford.
- Marston, D., Deno, S. L., & Tindal, G. (1983). *A comparison of standardized achievement tests and direct measurement techniques in measuring pupil progress* (Research Rep. No. 126). Minneapolis: University of Minnesota, Institute for Research on Learning Disabilities.
- Marston, D., & Magnusson, D. (1985). Implementing curriculum-based measurement in special and regular education settings. *Exceptional Children, 52*, 266-276.
- O'Connor, E. F. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research, 42*, 73-97.
- Parker, R. I., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reading measure: Construction methods, reliability, and validity. *The Journal of Special Education, 26*, 195-218.
- Parker, R. I., Tindal, G., & Hasbrouck, J. E. (1989). Initial validation of two classroom-based measures of reading comprehension. *Diagnostique, 14*, 222-240.
- Phillips, N. B., Hamlett, C. L., Fuchs, L. S., & Fuchs, D. (1993). Combining classwide curriculum-based measurement and peer tutoring to help general educators provide adaptive education. *Learning Disabilities Research & Practice, 8*, 148-156.
- Robinson, S. L., DePascale, C., & Roberts, F. (1989). Computer-delivered feedback in group-based instruction: Effects for learning disabled students in mathematics. *Learning Disabilities Focus, 5*, 28-35.

- Shin, J. (1999). *Reading-skill development and instructional practices facilitating reading growth for students with and without learning disabilities: A 1-year longitudinal study*. Unpublished doctoral dissertation, University of Minnesota, Minneapolis.
- Shin, J., Espin, C., Deno, S. L., & McConnell, S. (1999). *Technical characteristics of curriculum-based measurement in examining academic skill development and factors facilitating student growth in special education*. Unpublished manuscript, University of Minnesota.
- Shin, J., Deno, S. L., Robinson, S. L., & Marston, D. (in press). Predicting classroom achievement from active responding on a computer-based groupware system. *Remedial and Special Education, 21*, 53-60.
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.
- Willet, J. B. (1989). Questions and answers in the measurement of change. *Review of Research in Education, 15*, 345-421.
- Willet, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin, 116*, 363-381.