

## RESEARCH DESIGN AND METHODOLOGY SECTION

### Program Evaluation Using Hierarchical Linear Modeling with Curriculum-Based Measurement Reading Probes

Scott A. Stage  
*University of Washington*

This program evaluation used hierarchical linear modeling (HLM) to evaluate 99 ethnically diverse second-graders' growth in oral reading fluency using curriculum-based measurement (CBM) over the course of the school year. Five statistical advantages of the use of HLM are described and four are highlighted in this example: (a) The potential for improved accuracy in slope estimation of the empirical Bayes method used in HLM; (b) HLM provided the flexibility to test the effects of independent variables on the initial status of the slope and the change in slope over time; (c) HLM provided an analysis of the nested effects of the classroom on individual student performance; and (d) HLM provided a statistical test for both group effects and individual variation in reading growth. Results showed first-grade reading performance significantly predicted initial second-grade reading performance. In contrast, average classroom performance did not account for any differences in students' individual reading growth. Individual growth curves of 28 students who attended summer school showed that all of them made statistically significant growth in reading ability. A total of 71% of the summer school students scored between the first and second quartile of the spring, norming distribution at the end of summer school. This example shows that the use of HLM with normative CBM reading probes provided additional information not available with other statistical techniques.

This evaluation was conducted to demonstrate the use of hierarchical linear modeling (HLM) (Bryk, Raudenbush, & Congdon, 1996) with normative cur-

---

This evaluation was conducted for the Student Responsive Service Delivery system, which is supported by the Washington State Association of School Psychologists and the Washington State Office of Public Instruction.

The author thanks Harrah Elementary School and Jodi Sheppard, the school psychologist, for their forward thinking in the support of all students' opportunity to learn.

Address correspondence to Scott A. Stage, University of Washington, 322-R Miller, Box 353600, Seattle, WA 98195. E-mail: sstage@u.washington.edu.

riculum-based measurement (CBM) reading probes. Examples of HLM in education research investigating growth over time include the following studies: (a) a longitudinal study that showed low-ability readers did not differ from learning disabled readers in the rate of reading growth (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996); (b) a study of reading growth for poor readers in the first-grade showed that verbal IQ–achievement discrepancy standard scores did not predict reading growth or lack of it (Stage, Abbott, Jenkins, & Berninger, 2001); and (c) a study that demonstrated that increases in reading growth were best facilitated with complex letter-sound instruction (Hart, Berninger, & Abbott, 1997). Although these examples provide insight into the use of HLM for educational research questions pertaining to learning rate, this example using HLM with CBM reading probes provides an illustration of how HLM, which explicitly evaluates slope, is well-suited for program evaluation in schools using CBM. Because CBMs show adequate concurrent validity and are sensitive to student progress over time (Deno, 1993; Marston, 1989; Tindal, 1993), scholars have advocated their use in monitoring student progress for enhancing instructional programming (Fuchs, 1989, 1993; Marston & Tindal, 1995; Shapiro, 1996a; Shinn, 1995), in developing local norms to make educational programming decisions (Habedank, 1995; Shinn, 1988), in improving assessment practices for minority students (Baker, Plasencia-Peinado, & Lezcano-Lytle, 1998; Shinn, Collins, & Gallagher, 1998), and in providing an assessment system to monitor inclusive service delivery for all students (Tilly & Grimes, 1998).

#### ADVANTAGES OF HLM

The following paragraphs outline some unique advantages of HLM over other statistical techniques. First, HLM models growth over repeated measures, yielding slope as an outcomes measure at the individual student level, using an empirical Bayes estimate. The empirical Bayes estimate uses the composite of the sample's slope estimate and the predicted value of individual's slope estimate. Thus, the empirical Bayes estimate weights an individual's slope by the explained variance of the group estimate of slope (Bryk & Raudenbush, 1992). This statistical technique takes into account the group estimate of trend, whereas multiple regression uses the ordinary least-squares method to calculate the individual's slope without regard to the group (Bryk & Raudenbush, 1992). In addition, HLM provides a measure of reliability or the percentage of the total variance around each parameter that is estimated in the model selected (Arnold, 1992). In HLM models where the reliability estimates are low, the empirical Bayes estimation procedure better explains slope than ordinary least-squares estimates of slope used in regression statistical models (Bryk & Raudenbush, 1992). The comparison of accuracy in the prediction of CBM scores is an important issue because accurate prediction of students' future academic progress is required to make educational program decisions. In comparing the

split-middle technique, which splits the data in half and then uses the median of each half of the data series to calculate slope (Shapiro, 1996b) with the ordinary least-squares method for determining slope, research indicates that the ordinary least-squares method is more accurate (Good & Shinn, 1990; Shinn, Good, & Stein, 1989). However, in an example that compared the empirical Bayes prediction with ordinary least-squares prediction in children's vocabulary growth, results showed that the error in prediction was greater for the ordinary least squares predictions (Bryk & Raudenbush, 1992). In the current evaluation, the two techniques were compared for the level of accuracy in predicting students CBM reading scores.

Second, HLM analyzes growth over time by fitting the slope to the individual level. The unit of measurement for HLM is each participant's initial y-intercept and slope. In contrast, repeated measures analysis of variance (ANOVA) tests the interaction of repeated measurements with participants' performance (Bryk & Raudenbush, 1992). In HLM, the individual students' slopes over measurement intervals are fitted to the group's average slope using the empirical Bayes statistic. In comparison, repeated measures, one-way ANOVA uses the interaction between participants and repeated measurement as the error term. The *F*-test in a one-way, repeated measures ANOVA is the mean square of repeated measures divided by the mean square of the individual differences about the repeated measures (Tabachnick & Fidell, 1987). The repeated measures, one-way ANOVA tests for the amount of variance explained by systematic repeated measurement that does not capture individual growth but the variance explained by the mean square of the participant by measurement interaction, whereas HLM uses each participant's slope as the unit of analysis. In addition, repeated measures ANOVA requires that each individual is measured at each measurement point. Missing data are not allowed in this type of analysis. However, HLM allows for missing data because the empirical Bayes procedure estimates the slope using the group data.

Third, multiple regression can provide a valid statistical test at the person level of analysis, if slope is used as the dependent measure. A person level of analysis refers to a unit of measurement of personal characteristics and not higher-order organizational variables. In this case, the researcher performs an ordinary least-squares analysis to determine each student's slope before conducting the multiple-regression analysis. The students' slopes are used as the dependent measure and the independent variables are entered into the equation to test their relationship to the students' slopes. However, in this type of analysis, the results only indicate the effect on students' average slope. An advantage of HLM is that it allows the researcher the flexibility to specify where the variable will affect the slope. The variable might be expected to influence the initial status of the slope at the y-intercept. For example, an important consideration in evaluating students' reading growth is accounting for their previous reading ability. Research suggests that students who are better readers tend to gain more in their reading ability than poor readers gain (Stanovich, 1986). Thus, students

who read better before the implementation of a reading program would be expected initially to read more efficiently than students who do not read as well. In contrast, using multiple regression with slope as a dependent measure would restrict the influence of the regressed variable to predict average growth and not initial status.

Fourth, HLM provides a third level of analysis for nested effects because of the organizational structure of the school setting. A nested effect occurs when the effect of an independent variable is confined to a subset of the participants (Tabachnick & Fidell, 1989). For instance, students are instructed in separate classrooms, and as such, each classroom setting provides a unique setting different from other classroom settings. Each student is nested within a classroom and does not experience the instruction provided in another classroom. HLM provides an advantage over multiple-regression analysis because multiple regression does not allow for statistical analysis of nested variables (Arnold, 1992). Researchers using multiple regression must aggregate or disaggregate data to accommodate nested hierarchical structures. Disaggregation requires the researcher to use a separate analysis for each classroom analyzed. Disaggregation is problematic because it reduces the number of participants and the power of the statistical test. In the case of aggregation, the researcher treats classroom variables as personal characteristic variables, disregarding the nested structure of students within classrooms. Aggregation affects the standard error and can lead to type I errors where the researcher falsely reports a statistical significance when one does not exist (Cronbach & Webb, 1975). In a large-scale study of student attitudes and learning environments, Wong, Young, and Fraser (1997) used both HLM and multiple-regression standardized coefficients, in part, to determine the difference in statistical findings with these two approaches. The HLM analysis used a nested design to determine the effect at the student level and then the effect of the classroom on the student's performance. The multiple-regression analysis disregarded the nested effect. The results showed 12 significant coefficients for the HLM analysis and 15 significant coefficients for the multiple regression analysis: a 20% difference in the number of statistical significant coefficients in favor of the multiple-regression analyses, suggesting a potential for inflated statistical significance.

Reading research using the multiple-regression, aggregated data approach indicated that students' reading achievement was affected by individual differences in addition to the reading ability of their classroom peers (Share, Jorm, Maclean, & Matthews, 1984). This research suggests that students who were instructed in classrooms with peers who read well made more reading progress than students who were instructed with less able readers. This is not a trivial issue in the use of CBM for program evaluation and educational programming decisions because aggregating student performance across classrooms to assess the appropriateness of a student's relative standing in a grade level may place the student at a disadvantage because of instructional methods used in the classroom. Students are instructed in a unique classroom environment and the stan-

standard to which they should be compared is the average achieving student in their classroom. Thus, the classroom effect on the slope in academic achievement should be determined for the different classrooms in which students are instructed (as discussed by Fuchs, 1998). Therefore, the third level of the HLM used in the present evaluation was the CBM reading average for each of the classrooms where students received instruction.

Fifth, HLM provides a statistical test for both group and individual differences across repeated measures. The group differences are tested with the empirical Bayes estimation technique using the combined data from all of the participants. The group test uses the students' slopes across measurement intervals, and a *t*-test is conducted to determine whether the average fitted slope is greater than the standard error of slopes (Arnold, 1992). The test of individual differences about the group estimate is determined by testing for the variability of the individuals' slopes about the group estimate. A  $\chi^2$  statistical test is used to determine whether there is significant individual variability in the slopes about the average fitted slope (Arnold, 1992). The integration of both single-case and group comparison models provides an advantage not illustrated by other statistical models (Nugent, 1996). Group comparison evaluations are criticized because significant findings are based on the average differences between groups or across measurement periods (White, 1984). In program evaluation using group comparison, there is overlap between the treated and control groups, indicating that many students do not benefit from the program even though statistical significance is often found. The weakness of the single-case design is a lack of generalization to other students (Kazdin, 1982). HLM provides a group statistical test and a statistical test for individual differences.

### PURPOSE AND RATIONALE FOR THIS PROGRAM EVALUATION

In addition to conducting this study to elucidate the benefits of using HLM with CBM, this evaluation was conducted to determine the answers to three questions. First, the CBM research literature does not provide information on the predictive validity of spring first-grade oral reading on fall second-grade oral reading fluency. The determination of the magnitude and reliability of spring first-grade oral reading fluency scores to predict fall second-grade oral reading fluency scores allowed school personnel the opportunity to know whether first-grade screening provided useful information in planning for early academic interventions for these students. Second, no other studies have systematically evaluated whether classroom differences in the students' reading ability effected the students' growth in oral reading fluency over the course of the school year (Share et al., 1984). In the case that students' reading growth was isolated to specific classrooms, identification of teaching practices and student ability would be further evaluated. Third, after identifying the students across the school year who consistently scored at or below the first quartile on the CBM test probes, the

effects of a summer school reading program were assessed to determine whether the students benefitted from the instruction they received.

### HIERARCHICAL EFFECTS ANALYZED IN THIS STUDY

Three analyses were conducted in this study. First, students' slopes in CBM reading fluency probes were used to predict their end-of-year CBM reading fluency score. Two different estimates of slope were used: (a) the empirical Bayes method used in HLM and (b) the ordinary least-squares method used in multiple regression. The accuracy of these predictions was tested using the absolute mean difference. Second, a three-level HLM was conducted using student change in slope at Level 1 as an outcomes measure. Students' spring first-grade CBM reading fluency was used to predict their initial second-grade CBM reading fluency at Level 2; the nested effect of classroom average CBM reading fluency on the slope of students' reading fluency was tested at Level 3. The HLM nomenclature (taken from Bryk et al., 1996) for this model follows.

The Level 1 model tested the effect of student change in second grade CBM scores over time:

$$Y = \pi_0 + \pi_1 * (\text{Time}) + E$$

Because HLM provides a hierarchical structure of analysis, each tested effect is presented as a regression equation with an intercept, a slope for each variable, and an error term. The Level 1 model indicates that students performance on CBM ( $Y$ ) was affected by the initial status or the y-intercept ( $\pi_0$ ) at Time 0. In addition, student performance was affected by the slope in CBM performance over four repeated measurements ( $\pi_1 * \text{Time}$ ) from Time0 to Time3. Time0 to Time3 are the times that the oral reading fluency probes were administered, namely, October, January, March, and May. A random error term ( $E$ ) was specified, indicating that students individual differences in intercept and slope were allowed to vary. The Level 1 model tested students initial CBM reading fluency and slope in CBM reading fluency over time.

The Level 2 Model tested the effect of first grade CBM scores on initial second grade CBM scores:

$$\pi_0 = \beta_{00} + \beta_{01} * (\text{First-grade CBM}) + R_0$$

$$\pi_1 = \beta_{10} + R_1$$

The Level 2 model tested the effect of students' first-grade spring CBM performance ( $\beta_{01} * \text{First-grade CBM}$ ) on the students' initial second-grade CBM performance ( $\pi_0$ ), taking into account the mean initial status ( $\beta_{00}$ ) and random error ( $R_0$ ). It should be noted that only students' spring first-grade CBM performance was used rather than using the slope of their CBM performance across the school year. The effect of students' spring first-grade CBM performance on

slope in second-grade CBM performance was not specified ( $\pi_1 = \beta_{10} + R_1$ ), indicating that it was allowed to vary randomly. The effect of first-grade oral reading fluency on the slope of second-grade oral reading fluency was not tested, the sole contribution of second-grade classroom reading performance on the individual's slope in reading performance could be tested at Level 3, which is explained below. It was believed that the use of the classroom variable would be a more authentic measure of the instructional influence on reading growth rather than the students' previous first-grade reading progress. The proximal effect of first-grade reading performance on the students' initial second-grade reading performance was of primary interest at this level of analysis because it spanned a time period where no instruction took place for the students. Thus, the Level 2 model tested the effect of students' spring first-grade CBM performance on their initial second-grade CBM performance.

The Level 3 model tested the effect of classroom average CBM scores on second grade CBM slope:

$$\beta_{00} = \gamma_{000} + U_0$$

$$\beta_{01} = \gamma_{010} + U_1$$

$$\beta_{10} = \gamma_{100} + \gamma_{101} * (\text{Class Avg. CBM}) + U_2$$

The Level 3 model tested the effect of the students' second-grade classroom performance on their individual reading growth over the school year. Again, the error term (U) was allowed to vary randomly. The nested effect of average classroom CBM performance ( $\beta_{10} = \gamma_{100} + \gamma_{101} * \text{Class Avg. CBM} + U_2$ ) on individual students' slope in second-grade CBM performance ( $\pi_1 = \beta_{10} + R_1$ ) was analyzed at this level. The model explicitly tested the effect of classmates' average reading fluency on individual student's slope in CBM performance over time. The Level 3 analysis determined whether the students' slope in CBM performance was affected by the level of performance of their classroom peers.

In addition to the three-level HLM analysis described above, a separate HLM analysis was conducted to determine the increase in the summer school students' slopes in CBM performance over the course of the school year. This analysis was identical to the Level 1 analysis tested in the previous model, although it was conducted only with students who attended summer school. It was conducted as a separate analysis because the students displayed a negative effect on the overall second-grade CBM slope. Their poor growth in CBM oral reading fluency was used as one of the criteria to enroll them in summer school. The rationale for the analysis was to test whether summer school students gained in CBM reading performance as a group, compared with general education students. In addition, this model tested whether there was individual variability in reading growth among the summer school students. The result of the group test of the summer school students was used to represent graphically the difference in their growth compared to all the second-grade students' growth.

## METHOD

### Participants

The students who participated in this evaluation attended an elementary school in a rural agricultural area that was part of a larger study evaluating the effects of a general education initiative to provide educational services for all students experiencing academic difficulties. A total of 81% of the student population were eligible for free or price-reduced lunch. The school population was 15% European American, 20% Hispanic, and 65% Native American. The entire second grade participated in this program evaluation. There were 99 students: 50 males and 49 females. One student was African American, 16 were European American, 23 were Hispanic, and 59 were Native American. None of the students had been retained and none received special education services. Students were instructed in four different classrooms with a similar number of students in each class. A total of 28 of the 99 students attended summer school (17 males and 11 females). The summer school students were selected because they consistently performed below the 25th percentile on the CBM testing sessions. Four of these students were European American, 16 were Hispanic, and eight were Native American. A  $\chi^2$  statistical test was performed to determine the association between gender and attendance in summer school. The results showed no statistical significance of gender being associated with enrollment in summer school ( $\chi^2[1, N = 99] = 1.62; p = .20$ ). Another  $\chi^2$  statistical test was performed to determine the association between minority status and enrollment in summer school. Racial group was associated with educational placement in summer school ( $\chi^2[2, N = 99] = 26.1; p < .001$ ). A cross-tabs frequency table showed Hispanic students were overrepresented in summer school ( $n = 16$ ), compared with only general education enrollment over the nine-month school year ( $n = 7$ ).

### Procedure

CBM reading fluency norms were developed for the entire elementary school. Each grade level used unique reading passages taken from the reading curriculum appropriate for that grade level. The second-grade reading curriculum was published by Open Court Reading (1995). Passages used only for norming procedures were selected for each grade level. Teachers agreed not to use them during their instruction so students would not be exposed to these passages before CBM testing. The assessment procedures followed the procedures described in Shinn's (1989, pp. 239) *Curriculum-Based Measurement* text. Each student was given three 200-word passages and three 1-minute timed reading trials. Only the number of correct words read was recorded. The median score was recorded for each student. The norming process used 14 assessors: teacher assistants, parents, and retired teachers who were trained and supervised by the school psychologist. The timed readings were conducted in the gymnasium. A schedule that indicated when each classroom would be tested was distributed the week of testing. A su-

pervising teacher monitored students who waited in line for the next available assessor. After the students completed their timed readings, they brought the results to a desk at the front of the gym. The school psychologist entered the students' results in an Excel spreadsheet for later analysis. Testing of all students was completed in a day on each norming date. CBM norming occurred at the end of October, January, March, and May. Thirteen students missed one of the testing sessions. Classroom results were graphed using a box plot that identified the normal range and percentile ranks for each student by grade level. At the end of the school year, teachers nominated students for summer school based on end-of-year CBM performance and reading achievement on classroom assignments. CBM test results indicate that nominated students scored below the 25th percentile rank.

Students who attended summer school were instructed on a half-day schedule. Instruction was provided with an enriched reading curriculum that allowed for more one-on-one reading instruction and small group instruction accompanied by CBM oral reading fluency progress monitoring. On average, students attended 25.5 summer school days with a standard deviation of 4.4 days. The range was from 16 days to 29 days.

## RESULTS

### Accuracy of Prediction Comparing Empirical Bayes and Ordinary Least Squares

Table 1 shows each student's actual median score on the number of correct words read per minute in May, the empirical Bayes prediction of this score, and the ordinary least-squares prediction of this score using October, January, and March CBM scores. The next two columns in Table 1 show the empirical Bayes and ordinary least-squares error in the prediction of the May score. The overall mean score in the number of correct words read in May was 66.78. The overall mean for the empirical Bayes and ordinary least-squares predictions was 67.11 for both prediction techniques. The difference in the mean error of prediction using absolute numerical values for the empirical Bayes and ordinary least-squares predictions was 9.94 and 10.96, respectively. The difference in the mean error of prediction was not statistically significant ( $t[98] = -1.28; p = .20$ ). These results indicate that the absolute value in the error of prediction was about 10 words and 11 words for the prediction techniques, respectively. However, on average the difference in prediction using either slope prediction technique with the students' actual score was less than one word. This indicates that both techniques were comparably accurate and that on average they were quite accurate.

For further interpretation of this finding, the HLM reliability estimate of slope for CBM performance was .97. The reliability estimate is the amount of variance explained by the parameter estimate. This indicates that the amount of variance explained by the empirical Bayes estimate of students' ordinary least-squares

TABLE 1. Comparison of Hierarchical Model Predictions with Ordinary Least Squares of Words Read per Minute at the End of the School Year

Student	WPM	Empirical Bayes Prediction	OLS Prediction	Empirical Bayes Error	OLS Error
1	26.00	42.04	34.00	16.04	8.00
2	69.00	59.86	70.00	-9.14	1.00
3	82.00	96.50	118.33	14.50	36.33
4	74.00	82.25	70.67	8.25	-3.33
5	66.00	70.36	82.67	4.36	16.67
6	85.00	85.23	94.00	.23	9.00
7	45.00	53.34	47.25	8.34	2.25
8	50.00	50.72	47.00	.72	-3.00
9	67.00	82.83	76.00	15.83	9.00
10	59.00	71.99	89.00	12.99	30.00
11	123.00	102.83	111.33	-20.17	-11.67
12	62.00	61.10	64.33	-.90	2.33
13	70.00	84.32	87.67	14.32	17.67
14	74.00	74.55	88.00	.55	14.00
15	21.00	31.47	24.00	10.47	3.00
16	46.00	50.21	52.67	4.21	6.67
17	64.00	56.33	66.67	-7.67	2.67
18	19.00	29.96	19.50	10.96	.50
19	47.00	48.89	48.00	1.89	1.00
20	86.00	92.63	101.00	6.63	15.00
21	93.00	83.17	77.00	-9.83	-16.00
22	29.00	47.14	45.67	18.14	16.67
23	35.00	40.15	37.00	5.15	2.00
24	31.00	39.24	30.67	8.24	-.33
25	77.00	64.16	71.33	-12.84	-5.67
26	76.00	92.13	93.00	16.13	17.00
27	101.00	85.41	87.33	-15.59	-13.67
28	129.00	87.02	81.33	-41.98	-47.67
29	153.00	133.62	135.67	-19.38	-17.33
30	68.00	74.10	71.50	6.10	3.50
31	76.00	45.38	44.00	-30.62	-32.00
32	66.00	79.26	61.00	13.26	-5.00
33	37.00	52.66	55.67	15.66	18.67
34	112.00	119.55	104.75	7.55	-7.25
35	72.00	46.22	53.00	-25.78	-19.00
36	145.00	153.31	169.33	8.31	24.33
37	51.00	63.77	72.33	12.77	21.33
38	65.00	87.80	106.00	22.80	41.00
39	58.00	54.94	64.67	-3.06	6.67
40	36.00	36.75	29.00	.75	-7.00
41	47.00	46.38	47.00	-.62	.00
42	36.00	40.13	37.67	4.13	1.67
43	36.00	32.71	31.33	-3.29	-4.67
44	171.00	167.55	164.33	-3.45	-6.67
45	125.00	127.80	137.00	2.80	12.00
46	38.00	43.14	46.67	5.14	8.67
47	63.00	56.39	64.67	-6.61	1.67
48	50.00	50.99	63.67	.99	13.67
49	57.00	59.46	58.00	2.46	1.00
50	69.00	73.08	88.67	4.08	19.67
51	146.00	157.69	141.33	11.69	-4.67
52	67.00	66.16	77.33	-.84	10.33
53	75.00	79.47	66.67	4.47	-8.33

TABLE 1. *Continued*

Student	WPM	Empirical Bayes Prediction	OLS Prediction	Empirical Bayes Error	OLS Error
54	61.00	66.62	60.67	5.62	-.33
55	71.00	67.00	73.33	-4.00	2.33
56	108.00	106.70	102.00	-1.30	-6.00
57	63.00	69.93	75.25	6.93	12.25
58	78.00	60.74	64.00	-17.26	-14.00
59	12.00	22.17	7.00	10.17	-5.00
60	28.00	37.37	33.00	9.37	5.00
61	139.00	122.50	106.00	-16.50	-33.00
62	13.00	23.81	13.33	10.81	.33
63	100.00	84.93	91.67	-15.07	-8.33
64	83.00	85.19	95.33	2.19	12.33
65	106.00	126.56	142.67	20.56	36.67
66	130.00	95.08	104.00	-34.92	-26.00
67	40.00	60.32	66.33	20.32	26.33
68	87.00	79.39	96.00	-7.61	9.00
69	16.00	29.13	20.25	13.13	4.25
70	95.00	87.05	93.67	-7.95	-1.33
71	21.00	24.12	15.00	3.12	-6.00
72	37.00	42.64	38.67	5.64	1.67
73	23.00	35.92	20.00	12.92	-3.00
74	72.00	85.61	67.00	13.61	-5.00
75	82.00	58.40	50.50	-23.60	-31.50
76	42.00	38.76	35.00	-3.24	-7.00
77	39.00	40.50	37.33	1.50	-1.67
78	65.00	75.81	94.67	10.81	29.67
79	74.00	83.52	90.33	9.52	16.33
80	25.00	31.89	21.67	6.89	-3.33
81	56.00	53.39	45.25	-2.61	-10.75
82	21.00	33.93	26.33	12.93	5.33
83	37.00	46.35	48.33	9.35	11.33
84	26.00	32.12	26.67	6.12	.67
85	19.00	23.53	10.33	4.53	-8.67
86	48.00	39.50	34.33	-8.50	-13.67
87	93.00	109.34	111.33	16.34	18.33
88	13.00	25.18	16.00	12.18	3.00
89	75.00	86.76	77.67	11.76	2.67
90	77.00	86.98	96.33	9.97	19.33
91	103.00	99.26	110.00	-3.74	7.00
92	65.00	56.88	59.67	-8.12	-5.33
93	103.00	109.49	106.00	6.49	3.00
94	131.00	156.25	154.33	25.25	23.33
95	82.00	87.78	80.00	5.78	-2.00
96	37.00	38.20	25.75	1.20	-11.25
97	26.00	42.06	33.33	16.06	7.33
98	73.00	59.19	68.00	-13.81	-5.00
99	112.00	109.63	100.67	-2.37	-11.33
Mean	66.78	67.11	67.11		
SD	35.30	33.69	34.95		
Mean error using absolute scores				9.94	10.96
Range				-41 to 25	-47 to 36

*Note.* OLS = ordinary least squares; WPM = words per minute. WPM was the student's actual May score. Both the empirical Bayes and OLS predictions were based on WPM obtained in October, January, and March.

slope in oral reading fluency was quite high. Thus, the empirical Bayes estimation procedure did not strongly influence the fitted slope because the majority of the variance for the individual students' slopes did not differ much from the ordinary least-squares estimate.

### The Effect of First-Grade Reading Performance on the Initial Status of Second-Grade Reading Performance and the Effect of Classroom Reading Average on Second-Grade Reading Growth

Table 2 shows the results of the HLM fixed-effects analysis that modeled the students' mean reading growth over the four CBM norming dates: October, January, March, and May. It is considered a fixed-effects analysis because the parameters specified in the model are fixed by the model specified by the researcher. The y-intercept or initial status of the students' slopes was determined in combination with first-grade reading fluency as measured by the number of correctly read words per minute (WPM) in May of the students' first-grade school year. The initial status coefficient is the average median WPM read in October (i.e., 3.703). The coefficient shown with the first-grade WPM is the standardized contribution to the initial status (First-grade WPM  $\times$  0.849). For example, the median performance was 20 WPM for the first-grade May CBM probe. The calculation for students at the median would show that their initial status would be 21 WPM [ $3.703 \text{ WPM} + (\text{First-grade } 20 \text{ WPM}) \times 0.849 = 20.68 \approx 21$ ]. The initial status (i.e., 3.703) in the fixed effect model did not contribute significantly to the model prediction of initial second-grade reading status. However, the students' word reading fluency measured at the end of first grade did contribute significantly to the initial status of their second-grade reading fluency (First-grade WPM coefficient = 0.849;  $p < .0001$ ). This indicates that the students' end-of-year, first-grade reading fluency strongly predicted their initial second-grade reading fluency.

TABLE 2. Effects of First-Grade CBM Reading Fluency<sup>a</sup> and Classroom Reading Average<sup>b</sup> on Second Grade CBM Reading Fluency

Fixed Effect	Coefficient	Standard Error	<i>t</i> Ratio	<i>p</i> Value
Words read per minute				
Initial Status	3.703	1.78	2.08	0.121
First-grade WPM <sup>c</sup>	0.849	0.05	18.39	<0.0001
Slope in words read per minute				
Intercept	12.394	0.76	16.37	<0.0001
Second-grade class WPM	0.364	0.30	1.20	0.353

Note.  $N = 99$ . <sup>a</sup>First-grade reading fluency was based on the number of correctly read words in May. <sup>b</sup>Classroom reading average was determined by taking the average CBM reading probe performance over October, January, March, and May. There were four classrooms and the average was 36.25. Classes ranged from 33 to 40 words correctly read per minute. <sup>c</sup>WPM = words correctly read per minute.

The next two coefficient values show the change in slope and the contribution of the classroom average WPM to the prediction of the students' slope in reading across the months that the CBMs were administered. The results indicate that students grew significantly in reading fluency over the course of the year (slope in WPM = 12.394;  $p < .0001$ ), but that classroom average reading fluency did not have a significant effect (Second-grade class WPM coefficient = 0.364;  $p = .35$ ). The classroom averages were 32.6, 36.1, 36, and 39.6 WPM. The slight variability in classroom average performance suggests that, on average, the classroom reading differences were minimal and the usefulness of this variable to predict differences in slope would be expected to be weak. However, students did gain significantly in their reading fluency across the year.

Because first-grade reading fluency significantly predicted initial second-grade reading fluency, an illustration of the difference in initial level and slope in second-grade reading fluency is shown in Figure 1. A frequency distribution of first-grade CBM scores showed that WPM ranged from 0 to 127. The 25th percentile corresponded to 12 WPM; the 50th percentile corresponded to 21 WPM; and the 75th percentile corresponded to 46 WPM. To determine the initial second-grade status of a student who scored at the 25th percentile in first-grade, the initial status coefficient of 3.7 WPM was added to the product of the student's 12 WPM multiplied by the first grade coefficient of 0.849. This yielded 14 WPM

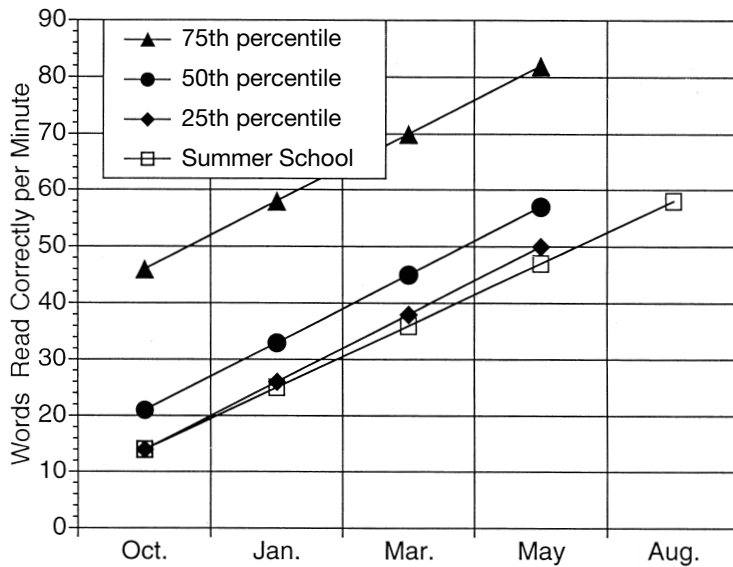


FIGURE 1. Second-grade reading growth by first-grade quartiles and summer school.

$(3.7 + [12 \times 0.849 = 10.19]) = 13.88 \approx 14$  WPM). The student who scored at the 25th percentile at the end of first-grade would be expected to read 14 WPM in the fall of second grade. In addition, the average slope was 12.39 WPM per measurement interval. Therefore, the student would be expected to read about 14, 26, 38, and 50 WPM at the CBM intervals of October, January, March, and May, respectively. A similar equation was used for a student who scored at the 50th and the 75th percentile rank. Figure 1 shows that the difference between the 25th and 50th percentile ranks was 7 WPM and the difference between the 50th percentile and the 75th percentile ranks was 25 WPM. The range in WPM scores between percentile ranks indicates that the distribution was positively skewed. Therefore, the HLM analysis was also conducted after transforming the first-grade reading fluency variable by taking the square root of each score (see Tabachnick & Fidell, 1989, for a discussion on the transformation of non-normal distribution scores). The results yielded similar findings. Because the statistical results were not altered, the results of the initial analysis are considered valid for interpretation and graphic representation.

### Average Reading Growth and Individual Differences for Students in Summer School

A separate analysis using only the data from the students who attended summer school is described next so that a comparison of their progress to the total second-grade sample can be examined. Table 3 shows the results of the fixed effect model of the mean summer school student reading growth and the random effect statistical analysis test for the individual variability in the summer school students' initial reading fluency and growth. The random-effects model differs

TABLE 3. Fixed and Random Effects of the Summer School Students Second-Grade CBM Reading Fluency

Fixed Effect	Coefficient	Standard Error	<i>t</i> Ratio	<i>p</i> Value	
Words Read per Minute					
Initial Status	14.111	2.43	5.80	<0.0001	
Slope in Words Read per Minute <sup>a</sup>					
Intercept	10.767	0.92	11.63	<0.0001	
Random Effect	Standard Deviation	Variance Component	<i>df</i>	$\chi^2$	<i>p</i> Value
Words Read per Minute					
Initial Status	11.939	142.544	27	193.732	<0.0001
Slope	4.487	20.140	27	168.345	<0.0001
Error	6.202	38.471			

*Note.* *N* = 28. The fixed-effect model tests the groups initial status and growth. The random-effect model tests for individual differences in initial status and growth. <sup>a</sup>CBM reading probes for summer school students were collected in October, January, March, May, and August. *df* = degrees of freedom.

from the fixed-effects model in that the random-effects model tests whether there are a significant number of students who varied from the slope estimated for the group. For the fixed-effect model, the mean initial status and slope in the students' reading fluency were both statistically significant ( $p < .0001$ ), meaning that the students' initial number of correctly read words in October significantly contributed to the prediction of reading growth over the course of school year. In addition, the summer school students' slope in word reading fluency from October to August significantly increased.

A graphic representation of the summer school students' reading growth is also shown in Figure 1. Their initial status in October was 14 WPM as noted by the initial status coefficient in Table 3. The slope coefficient indicated that summer school students' reading grew by 10.767 correctly read words per minute across the five CBM norming intervals. The graph shows the mean WPM in October, January, March, May, and August as 14 WPM, 25 WPM, 36 WPM, 47 WPM, and 58 WPM, respectively. Their final scores indicate that by the end of summer school, the students gained to a level commensurate with the spring reading level of the general education students who were identified in first-grade as reading at the 50th percentile rank. Those students read 57 WPM in May of their second-grade school year and the summer school students read 58 WPM in August of their second-grade school year.

Table 3 also shows the random-effect results that indicate the individual variability in the summer school students' initial reading status and slope in word reading fluency across their second-grade school year. The results show that there were significant individual differences in both initial status and growth in word reading fluency over the school year. The individual variability about the initial status for the summer school students indicates many of them did not conform to the average initial status determined for the group. Therefore, many summer school students started the school year reading fewer than 14 WPM and some may have read more than 14 WPM. In addition, the significant  $\chi^2$  for slope indicates that many of the students either gained less or more than the 10 WPM as was determined for the group. To investigate the number of individual students who made significant reading growth, the student's slope was divided by twice the standard error. The results indicated that all the students made significant growth at  $p < .05$ . A comparison of the summer school students' August level of WPM with the general education students' May level of WPM showed that 71% (20 of 28 students) of the summer school students read at least 50 or more WPM. This relatively high percentage of students reading 50 WPM indicates that the majority of the summer school students' word reading fluency was above the 25th percentile rank as determined by the second-grade May CBM distribution.

## DISCUSSION

Five advantages of the use of HLM were presented and four subsequent examples were tested in this program evaluation of second grade students' reading

growth using CBM over the school year. The first HLM advantage tested was the predictive accuracy of the empirical Bayes estimate compared with the ordinary least-squares estimate determined in bivariate and multiple-regression statistical analyses. The empirical Bayes estimate was slightly more accurate than the ordinary least-squares estimate although not at a statistically significant level. In fact, on average the difference in prediction of the students' end-of-year CBM word reading fluency using either of these statistical estimates of slope was within one word of the group's actual number of correctly read words per minute. The lack of predictive advantage of the empirical Bayes estimate over the ordinary least-squares procedure was likely because of the reliability in slope estimate parameter was .97. The high reliability of slope measurement indicates that linear estimate of slope over the three CBM measurement intervals was quite accurate so that the use of the empirical Bayes method provided little improvement over the ordinary least-squares method.

The second HLM advantage described was the flexible modeling procedures that allow the program evaluator to specify whether the selected variables will influence the initial status or whether the variables will influence the slope. Because research has shown that students' previous reading ability affects their future reading ability (Stanovich, 1986), students' first-grade word reading fluency was used to predict their initial second-grade reading ability. First-grade word reading fluency strongly predicted initial second-grade reading fluency. This information can be used to make instructional program decisions. In the case of the school that participated in this evaluation, this information was used to identify second-grade students who were likely to have continued reading difficulty in third grade. Students were nominated to attend summer school based partially on these results. For instance, students who scored below the 25th percentile on the end-of-year CBM testing were initially considered for summer school.

The fact that students' first-grade reading fluency predicted their second-grade reading fluency is an important contribution. Although there is ample research about the concurrent validity of CBM oral reading fluency probes (Good & Jefferson, 1998), the predictive validity of CBM oral reading fluency has not been confirmed. This strong predictive validity estimate suggests, at least for the transition from first to second grade, CBM reading fluency measures can be used to identify students who are apt to be academically at risk. In addition, the determination of the students' slope in oral reading fluency can be used to plan effectively for additional reading interventions used to increase students' reading fluency. For instance, the average second-grade student in this study started the school year reading 21 WPM and showed a slope of 12 WPM per quarter. If students start the academic year far below the initial level of 21 WPM, then they will have to learn more than 12 WPM per quarter to maintain pace with the average student. This result provides a statistically accurate means of evaluating students' academic progress and allows for educational programming decisions (Tilly & Grimes, 1998).

The third HLM advantage tested was the nested effect of students' classroom average word reading ability on reading growth. Research indicates that stu-

dents' reading progress is influenced by their classmates' ability level (Share et al., 1984). In cases where there are differences in the classroom-based slope in word reading fluency, evaluation of individual students' reading ability should be compared with this standard (Fuchs, 1998). In the HLM analysis reported in this evaluation, the effect of the average classroom word reading fluency did not contribute to the prediction in slope in word reading fluency across the school year. Although this finding would need to be replicated in other local norming CBM evaluations, it suggests that differences in students' ability nested within classrooms did not have an effect on students' reading progress. Typically, the effect of classrooms has not been taken into account in educational research (Arnold, 1992). However, as noted by Fuchs (1998), classroom is an important consideration because students should be compared with the standard student performance within their classroom. Aggregating data without investigation into classroom differences may bias educational programming decisions made about students' progress without accounting for classroom differences (Cronbach & Webb, 1976). However, the difference between the four classrooms in this evaluation was slight. Other classroom variables, such as the proportion of students receiving instructional modifications, would determine whether instruction in classrooms where reading instruction was individually tailored to students' instructional needs explained differences in students' slope in reading fluency. For example, in Marston's (1987–1988) program evaluation of students who received both general and special education programs, he found that placement in special education resulted in an increase in CBM slope for word reading fluency. He further evaluated general and special education teachers' ratings of the type of instruction that they used and found that special educators rated home interventions, cooperative learning, and homework more highly than general education teachers. A more statistically rigorous way to evaluate instructional modifications would be to use the proportion of instructional modifications in an HLM analysis to evaluate the impact that they have at the classroom level. This sort of analysis would allow valid measurement of the classroom effects of modified instruction within the classroom context.

The fourth HLM advantage tested in this evaluation was the effect of the summer school program on the students' growth in oral reading fluency as a group and also at the individual student level. The advantage highlighted with this analysis is that traditional program evaluations using group statistical comparisons actually measure differences in terms of the group's average performance and thereby neglect the individual (White, 1984). In contrast, the use of single-case designs does not generalize to the entire group (Kazdin, 1982). The evaluation of the 28 students who attended summer school showed that they made significant progress over the course of the school year. When individual students' slopes in word reading fluency were tested against the standard error, it was found that all of them made significant progress at the  $p < .05$  level. When their August WPM scores were compared with the general education students' May WPM scores, it was found that 71% of the summer school students were reading above the 25th percentile. Although this is considered a sig-

nificant finding, there are several potential threats to validity that make causal statements suspect. For example, when evaluating interrupted time series designs such as this one, the change in level of the summer school students' reading compared with that of general education students may merely be the result of discontinuation of measurement of the general education students in the summer (see Cook & Campbell, 1979, for further discussion of threats to internal validity). Simply stated, it might be that if the general education students' word reading fluency had been measured in August, they might have shown a 12 WPM increase, negating the appearance that the summer school students had actually improved compared with the general education students. In any case, if the predictive trend of students' end-of-year word reading fluency continues to predict strongly the following year's academic performance, then it would be safe to say that the students had not lost ground. More importantly, the analysis of the individual student reading slopes indicate that all of them made significant progress. However, the eight students who did not make adequate progress can be targeted for further intervention and progress monitoring at the individual student level.

The HLM advantage not explicitly tested in this evaluation was the comparison with repeated measures ANOVA. As indicated previously, ANOVA results indicate the amount of variance explained across repeated measures without determining the participants' slopes. In addition, ANOVA does not allow for missing data (Tabachnick & Fidell, 1987). In the CBM norming across the school year, 14 students missed one testing session. Therefore, their data would not have been used in a repeated measures ANOVA. These features limit the practical usefulness of repeated measures ANOVA statistical models in CBM or program evaluations.

The HLM computer program (Bryk et al., 1996) runs on Windows operating systems. Data management can use ASCII input, SYSTAT file input, SAS transport file input, and residual file output can be transferred to SPSS. The nomenclature used in the descriptions of the model presented in this article is the same as those presented on the computer screen. The evaluator simply specifies the parameter effects by clicking the mouse. The HLM computer program writes the equations using the format used in this article. The residual file provides some measures to be considered in the adequacy of the data set such as the Mahalanobis distance for the identification of outlier data points. The fixed variable, ordinary least squares, and the empirical Bayes measures of the y-intercept and slope data are provided for each participant in the residual file. However, the researcher must calculate each participant's slope using the fixed variable estimate and the preferred slope measurement of either the ordinary least-squares measure or the empirical Bayes measure. This calculation requires an additional step to determine the slope for each student. Another difficulty in using the HLM residual file is that the participants' raw data are not printed in this file. This limitation requires the researcher to aggregate these data files if further analysis using the data from both files is desired. Singer (1998) described the use of SAS

PROC MIXED, which performs similar multilevel analysis using SAS software packages.

Because the use of HLM statistical techniques is relatively new, there are unanswered questions about the adequacy of sample size used (Rogosa & Saner, 1995). In general, the number of cases to variables used in this evaluation provided a moderate sample size that should ameliorate errors in standard error estimation.

In summary, the four advantages tested in this example highlighted the flexible nature of HLM in determining initial effects on slope and factors affecting slope that are not readily available in other statistical techniques. In addition, the HLM advantage of providing statistical tests for both group and individual variation in slope offers a unique form of evaluation not displayed in other traditional program evaluation techniques. Future research will no doubt determine the statistical adequacy in HLM with evaluations conducted in nested structures such as those found in the organizational structure of schools. In addition, the fit between HLM analysis and evaluations using CBM can provide valuable information about the effects of educational interventions on students rate of learning.

One final contribution of this program evaluation and the advantage in the use of CBM is the fact that the majority of the students in this program evaluation were minority students of limited economic means (i.e., 23% Hispanic and 59% Native American; 80% receiving free or price-reduced lunch) who showed significant increases in their growth in oral reading fluency. A recent study found ethnic and gender bias in the use of CBM oral reading fluency scores measured at a single point in time (Kranzler, Miller, & Jordan, 1999). This finding indicates that the use of CBM in the identification of students may be biased. However, the purpose for using CBM in this evaluation was twofold: (a) CBM was used for screening and (b) CBM was used for progress monitoring (as described by Marston 1987–1988). The fact that 71% of the students who received summer school instruction increased their CBM score to the first quartile of the average general education student's spring reading score highlights the importance of using CBM slopes for progress monitoring rather than solely as an identification procedure leading to special education (as described by Shinn et al., 1998).

Although this article highlights the use of HLM with CBM, other uses of HLM in program evaluation include modeling the change in social behavior over time. For example, in a program demonstration project, students at risk for developing serious emotional disturbance are receiving positive behavioral support interventions to reduce disruptive school behaviors while increasing prosocial behaviors (Cheney, 1999). The Level 1 model of growth over time uses behavioral observation data collected in the students' classrooms on a weekly basis. The change in slope of disruptive behavior and academic engaged time will be used as outcome measures. The Level 2 model uses treatment dosage or the amount of intervention which is determined by the number of sessions the student receives to teach him or her how to behave effectively in the classroom. The Level 2 variables will include social skills training, contingent rewards for

prosocial behavior, and academic interventions for students with poor academic performance. The Level 3 model includes four different schools that are participating in the project. Therefore, HLM can be used for other types of program evaluations of interest to school psychologists besides CBM.

## REFERENCES

- Arnold, C. L. (1992). An introduction to hierarchical linear models. *Measurement and Evaluation in Counseling and Development*, 25, 58–90.
- Baker, S. K., Plasencia-Peinado, J., & Lezcano-Lytle, V. (1998). The use of curriculum based measurement with language-minority students. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 175–213). New York: Guilford.
- Bryk, A. S., Raudenbush, S. W., & Congdon, R. T., Jr. (1996). *Hierarchical linear and nonlinear modeling with HLM/2L and HLM/3L programs*. Chicago: Scientific Software International.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.
- Cheney, D. (1999). *Washington's assessment: An intervention program for students with emotional disturbance: Project WAIS-ED*. Funded by the US Office of Special Education and Rehabilitation Services, Washington, DC.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.
- Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude \* treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology*, 67, 717–724.
- Deno, S. L. (1993). Curriculum-based measurement. In J. J. Kramer (Vol. Ed.) & J. C. Conoley (Series Ed.), *Curriculum-based measurement* (pp. 1–23). Lincoln, NE: Buros Institute of Mental Measurements and Department of Educational Psychology, University of Nebraska–Lincoln.
- Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, 88, 3–17.
- Fuchs, L. S. (1989). Evaluating solutions, monitoring progress, and revising intervention plans. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 153–181). New York: Guilford.
- Fuchs, L. S. (1993). Enhancing instructional programming and student achievement with curriculum-based measurement. In J. J. Kramer (Vol. Ed.) & J. C. Conoley (Series Ed.), *Curriculum-based measurement* (pp. 65–103). Lincoln, NE: Buros Institute of Mental Measurements and Department of Educational Psychology, University of Nebraska–Lincoln.
- Fuchs, L. S. (1998). Computer applications to address implementation difficulties associated with curriculum-based measurement. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 89–112). New York: Guilford.
- Good, R. H., III, & 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., III, & Shinn, M. R. (1990). Forecasting accuracy of slope estimates for reading curriculum-based measurement: Empirical evidence. *Behavioral Assessment*, 12, 179–193.
- Habedank, L. (1995). Best practices in developing local norms for problem solving in the schools. In A. Thomas & J. Grimes (Eds), *Best practices in school psychology—III* (pp. 701–715). Washington DC: National Association of School Psychologists.
- Hart, T. M., Berninger, V. W., & Abbott, R. D. (1997). Comparison of teaching single or multiple orthographic-phonological connections for word recognition and spelling: Implications for instructional consultation. *School Psychology Review*, 26, 279–297.

- Kazdin, A. (1982). *Single case research designs*. New York: Oxford University Press.
- Kranzler, J. H., Miller, M. D., & Jordan L. (1999). An examination of racial/ethnic and gender bias on curriculum-based measurement of reading. *School Psychology Quarterly, 14*, 327–342.
- Marston, D. (1987–1988). The effectiveness of special education: A time series analysis of reading performance in regular and special education settings. *Journal of Special Education, 21*, 13–26.
- 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.
- Marston, D., & Tindal, G. (1995). Best practices in performance monitoring. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology—III* (pp. 597–636). Washington DC: National Association of School Psychologists.
- Nugent, W. R. (1996). Integrating single-case and group-comparison designs for evaluation research. *Journal of Applied Behavioral Science, 32*, 209–226.
- Open Court Reading. (1995). *Collections for young scholars*. Chicago and Peru, IL: SRA/McGraw-Hill.
- Rogosa, D., & Saner, H. (1995). Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics, 20*, 149–170.
- Shapiro, E. S. (1996a). *Academic skills problems: Direct assessment and intervention* (2nd ed.). New York: Guilford.
- Shapiro, E. S. (1996b). *Academic skills problems workbook*. New York: Guilford.
- Share, D. L., Jorm, A. F., Maclean, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Educational Psychology, 76*, 1309–1324.
- Shinn, M. R. (1988). Development of curriculum-based local norms for use in special education decision-making. *School Psychology Review, 17*, 61–80.
- Shinn, M. R. (Ed.) (1989). Identifying and defining academic problems: CBM screening and eligibility procedures. *Curriculum-based measurement: Assessing special children* (pp. 90–129). New York: Guilford.
- Shinn, M. R. (1995). Curriculum-based measurement and its use in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology—III* (pp. 547–567). Washington, DC: National Association of School Psychologists.
- Shinn, M. R., Collins, V. L., & Gallagher, S. (1998). Curriculum-based measurement and its use in a problem-solving model with students from minority backgrounds. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 143–174). New York: Guilford.
- Shinn, M. R., Good, R. H., III, & Stein, S. (1989). Summarizing trend in student achievement: A comparison of methods. *School Psychology Review, 18*, 356–370.
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth curves. *Journal of Educational and Behavioral Statistics, 23*, 323–355.
- Stage, S. A., Abbott, R. D., Jenkins, J. R., & Berninger, V. W. (2001). *Predicting response to early intervention using Verbal IQ, reading-related language abilities, attention ratings, and Verbal IQ–word reading discrepancy*. Manuscript submitted for publication.
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360–407.
- Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper & Row.
- Tilly, W. D., III, & Grimes, J. (1998). Curriculum-based measurement: One vehicle for systemic educational reform. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 32–60). New York: Guilford.
- Tindal, G. (1993). A review of curriculum-based procedures on nine-assessment components. In J. J. Kramer (Vol. Ed.) & J. C. Conoley (Series Ed.), *Curriculum-based measurement* (pp. 25–64). Lincoln, NE: Buros Institute of Mental Measurements and Department of Educational Psychology, University of Nebraska–Lincoln.

White, O. R. (1984). Selected issues in program evaluation: Arguments for the individual. *Advances in Special Education, 4*, 69–121.

Wong, A. F. L., Young, D. J., & Fraser, B. J. (1997). A multilevel analysis of learning environments and student attitudes. *Educational Psychology, 17*, 449–468.

Action Editor: Timothy Z. Keith

Acceptance Date: August 4, 2000