Journal of Applied School Psychology, 25:270–285, 2009

ISSN: 1537-7903 print / 1537-7911 online
DOI: 10.1080/15377900802487201

Routledge
Taylor & Francis Group

Evaluating the Effectiveness of a Reading Remediation Program in a Public School Setting

JANE DOWNING, JASON WILLIAMS, and E. WAYNE HOLDEN
RTI International, Research Triangle Park, North Carolina, USA

This article reports an evaluation of a reading remediation program delivered to 151 high-risk students in a public school setting. Implementation fidelity was assessed by the amount of time each child was exposed to the program. Despite receiving fewer than the recommended hours of instruction, analyses of standard achievement scores indicated significantly greater change than expected in reading comprehension, reading fluency, and word attack skills. Number of individual hours of instruction yielded the most conservative estimates of change across time. Findings are discussed within the context of conducting effectiveness evaluations of educational interventions and the need to monitor implementation fidelity at the individual level to best inform decision making and resource allocation.

KEYWORDS reading remediation intervention, evaluating effectiveness, implementation fidelity, dosage, public school setting

Children with learning disabilities and children at risk for failure in school because of other factors face significant reading challenges. Because reading serves as the major learning pathway to other subject areas, children who fall behind in reading are at risk of falling behind in other subject areas and at substantial risk for failure in school (Lyon, 1998). There is considerable

This article was accepted under the editorialship of Dr. Charles A. Maher.
The authors gratefully acknowledge the financial support to conduct this study through a Grant Agreement from the North Carolina GlaxoSmithKline Foundation to RTI International. The reading remediation program tested in this study was developed and delivered by The Hill Center, a private not-for-profit educational institution in Durham, NC, devoted to the needs of children with learning disabilities and their families. RTI International is a trade name of Research Triangle Institute.
Address correspondence to Jane Downing, RTI International, 3040 Cornwallis Road, PO Box 12194, Research Triangle Park, NC 27709. E-mail: jdowning@rti.org

evidence that children who are poor readers in the early years of elementary school are likely to continue to be poor readers when they reach high school (Snow, Burns, and Griffin, 1998; Torgesen et al., 2006). In addition, negative economic and emotional consequences often follow struggling readers into adulthood (Lyon, 2002).

While there is a growing body of knowledge about effective preventive strategies to reduce reading failure in at-risk children (Torgesen et al., 1999; Vellutino et al., 1996), there is little empirical information about the effectiveness of reading interventions for children identified as struggling readers in public school settings (Blachman et al., 2004; Erlbaum, Vaughn, Hughes, & Moody, 2000; National Institute of Child Health and Human Development [NICHD], 2000; Torgesen et al., 2006). Furthermore, existing evidence suggests that many children who receive remedial reading support in schools make little or no progress (Moody, Vaughn, Hughes, and Fischer, 2000; Snow et al., 1998; Torgesen et al., 2006). Thus, there is a critical need to identify effective reading interventions for struggling readers, particularly interventions that succeed in changing the growth trajectory of children at risk for reading and school failure.

Whether tackling the need for assisting struggling readers or struggling learners more generally, educators and related services personnel, such as school psychologists, face the challenge of identifying successful, evidence-based interventions to enhance the academic performance of struggling students. Given limited resources, other feasibility constraints, and the need for results, practitioners can benefit from access to objective and valid data to guide their decision making. The availability of such evidence hinges on studies of interventions that examine not only changes in relevant outcome measures but also measures of implementation fidelity. As Dusenbury, Brannigan, Mathea, and Hansen (2003) note, "the key to understanding how successful research can be translated into successful practice lies in understanding how programs and policies can be implemented so that quality is maintained and the programmatic objectives intended by program developers are achieved" (pp. 237–238).

Within the education field and particularly within research on the effectiveness of school-based interventions or programs aimed at improving student learning, there has been limited research on implementation fidelity (Melde, Esbensen, & Tusinski, 2006). Furthermore, as noted in a relatively recent review of research on implementation fidelity, definitions and measures of fidelity have not been consistent across studies (Dusenbury et al., 2003). The largely absent consideration of findings on implementation fidelity and the lack of consistency in measures of implementation fidelity present significant challenges to adequately assessing the effectiveness of school-based educational interventions. This is especially problematic because studies examining implementation fidelity generally indicate that poor implementation is likely to result in a loss of program effectiveness.

More recently, there has been growing attention to implementation fidelity measures in school-based programs and intervention research (see, for example, Melde et al., 2006; Mokrue, Elias, & Bry, 2005; Ruiz-Primo, 2006; Torgesen et al., 2006). One aspect of implementation fidelity in need of research consideration is exposure to the intervention or implementation dosage. While educational intervention developers typically prescribe the amount of exposure to or dosage of the intervention necessary to effect change, few educational intervention studies have examined the actual amount of exposure relative to the amount recommended. Furthermore, in instances where data on the extent of intervention exposure are collected at the individual level, these data are seldom used in analytic modeling for change over time (e.g., growth models) in outcome measures. However, whether growth models use individual- or aggregate-level measures of dosage can have important implications for interpreting findings on intervention effectiveness, which in turn may have implications for how educators allocate resources for and monitor the impact of interventions aimed at helping struggling learners.

This study contributes to the growing body of research on school-based reading interventions for young, elementary-school-aged children. Within a public school setting, the study tested the effectiveness of a reading remediation program for children at high risk for reading failure. A comprehensive, multimodal intervention strategy delivered by exceptional-children (EC) teachers was used, and it was evaluated with standardized measures. In addition, the study addresses implementation fidelity in two important ways. First, exposure to the intervention was measured for each participating student. Using this measure of implementation fidelity made it possible to examine the variance between recommended and received hours of instruction. Second, the analysis compared two different models for examining the relationship between intervention exposure and change in reading achievement. The first model used simple time coding (i.e., average hours of instruction for all students in a given grade level) as the measure of intervention exposure; the second model used individual time scores (i.e., actual number of hours of instruction each student received).

METHOD

Elementary School Selection

Eight elementary schools selected from an urban school district in a mid-size Southeastern city participated in the study. Five of the eight schools in the study included kindergarten through grade 5 and the remaining three schools included prekindergarten through grade 5. Among the schools, student enrollment ranged from 200 to 747 students, with an average of 577 students. On average, 49% of students enrolled in the schools were Black,

34% were White, and 14% were Hispanic. And, on average, 48% of students were eligible for free or reduced-price lunch (National Center for Education Statistics, 2007).

Participants

Over the 3-year study period (2003–2006), 23 EC teachers from the eight selected schools participated in the study, and a total of 151 first- through fifth-grade students participated in the study. The 151 student participants were diverse in terms of gender, disability, and other factors (see Table 1). Paralleling the racial/ethnic distribution across the participating schools, 55% of the students were Black, 30% were White, and 15% were Hispanic. A greater percentage of males (66%) than females (34%) participated in the study. Moreover, these children were clearly at risk of not being successful in school. More than 60% of students were eligible for free or reduced-price lunch, and 79% were identified as eligible to receive exceptional children services (the term *exceptional child* typically refers to children with special problems related to physical disabilities, sensory impairments, emotional disturbances, learning disabilities, or mental retardation), with 63% of these children having a learning disability and approximately 10% having attention deficit/hyperactivity disorder (ADHD). Among these children, 63% had repeated a grade.

Participating EC teachers identified students for enrollment in the intervention program; the reasons the participants were selected varied. The most common reason, indicated for 42% of students, was that the student was at risk of failing his or her state-level end-of-grade tests. Other common reasons for selection included being a student with a learning disability (23%), being at risk of failing a grade (22%), or having an IQ lower than 85 (23%).

Intervention

The intervention program is a research-based method to improve the reading skills of public school students who do not meet the state standards for reading proficiency. The program is adapted from a remediation methodology specifically created as a public school program in which struggling readers are taken out of their regular classrooms daily for short periods to receive specialized instruction. The implementation of this program in a public school setting represents an attempt to scale up a methodology currently used primarily in a private school setting. The program addresses the five essential components of a successful reading program, as put forth in the National Reading Panel Report of 2000: phonological awareness, phonics, fluency, vocabulary, and comprehension (NICHD, 2000). The curriculum can be implemented in a class period ranging from 45 minutes to 1 hour.

TABLE 1 Demographic Data for Participating Children

Measure	Participating children	
	Number	Percent[a]
Grade level at entry to remediation program		
First grade	9	6.6
Second grade	31	22.6
Third grade	60	43.8
Fourth grade	25	18.3
Fifth grade	12	8.8
Total	137	100.0
Race		
Black	75	54.7
White	41	29.9
Hispanic	21	15.3
Total	137	100.0
Gender		
Male	90	65.7
Female	47	34.3
Total	137	100.0
Eligible for free or reduced-price lunch		
Yes	84	61.3
No	53	38.7
Total	137	100.0
English as a second language		
Yes	19	13.9
No	118	86.1
Total	137	100.0
Identified to receive exceptional children services		
Yes	108	78.8
No	29	21.1
Total	137	100.0
Type of learning difference[b]		
Learning disability (LD)	68	63.0
Attention deficit hyperactivity disorder (ADHD)	11	10.2
LD and ADHD	1	0.9
Other	28	25.9
Total	108[c]	100.0
IQ		
Below average (below 85)	59	52.2
Average (85–115)	54	47.8
Total	113[d]	100.0
Repeated a grade		
Yes	86	62.8
No	51	37.2
Total	137	100.0

[a]Totals may not equal 100 due to rounding.

[b]Type of learning difference for which student was eligible to receive exceptional children services.

[c]Because only 108 students (79%) were identified to receive exceptional children services, only those students are included here.

[d]Typically, only students identified to receive exceptional children services take IQ tests; this was not the case for all children participating in this study. In addition, IQ score data are missing for four students identified to receive exceptional children services. Seventy-nine students took the Wechsler Intelligence Scale for Children–3rd Edition (WISC-III), 15 took the Developmental Activities Screening Inventory (DASI), 11 took the Wechsler Intelligence Scale for Children–4th Edition (WISC-IV), and another 6 took the Universal Nonverbal Intelligence Test (UNIT). One child took three additional assessments: Cognitive Assessment System (CAS), Wechsler Preschool and Primary Scale of Intelligence (WPPSI), and the Wechsler Preschool and Primary Scale of Intelligence–Revised (WPPSI-R).

While students work in small groups of four, each student has an individualized curriculum to provide instruction in areas where there are demonstrated skill deficits. Students are assessed and assigned a level of instruction and use workbooks and readers that target specific phonetic patterns taught at each level. Small units of information are presented sequentially and practiced daily until a set criterion is met for 3 to 5 consecutive days and overlearning is achieved. Mastered skills are reviewed weekly to ensure retention. All student responses are graphed and charted daily by the teachers and students to document mastery before students advance to a higher skill level. Student–teacher interaction focuses on praise and positive reinforcement of correct answers or approximations of the correct response.

Approximately 45% of the students participating during the study period were enrolled for 1 school year only, 42% were enrolled for 2 school years, and 13% were enrolled for 3 years. Among participating students, 45% remained enrolled in the program through the end of the study period. Of those who exited the program earlier than anticipated, 31% exited because they moved, approximately 23% exited because they graduated to middle school, and about 25% exited because the teacher left the school, too few students were enrolled, parents withdrew their child from the program, or a classroom teacher refused to allow his or her student to participate.

Students received an average of 56.01 hours of instruction between their Time 1 (pretest) and Time 2 (first posttest) assessments ($N = 137$), occurring over an average of 8.78 months, and an average of 59.56 instructional hours between their Time 2 (first posttest) and Time 3 (second posttest) assessments ($N = 71$), occurring over an average of 11.55 months. According to the intervention developer, the recommended average number of expected instructional hours is 144 hours per year (i.e., 4 hours per week); over 2 years the average expected number of instructional hours would double to 288 hours. Comparing recommended with received hours of instruction shows that students in this study received less than half of the average expected number of expected instructional hours.

Public School Teacher Training

Of the 23 EC teachers who implemented the intervention program, 22 were female. All of the participating teachers had completed a bachelor's degree; 14 teachers had completed a master's degree. The median total years of overall teaching experience was 15 years. The median total years of teaching at their current school was 2.5 years.

Over the 3-year study period, the EC teachers were trained to implement the intervention program. First, teachers participated in a 4-day training about 1 month prior to the beginning of the school year. Three days of training focused on the reading, written language, and math methodology used

in the private school setting—a precision teaching method using charting and graphing of student progress. One day of training focused on the intervention program, emphasizing the fundamentals of phonological awareness and phonics as a bridge to improving reading fluency and comprehension. The training included 45 hours of coursework and a minimum of five classroom observations with direct feedback over 18 months. In addition to the site visits, three 3-hour follow-up training meetings were conducted with all participating teachers at roughly 3, 6, and 8 months after the initial training. Approximately 1 year after their initial training, all participating teachers attended a 2-day follow-up training that included a review of the intervention methodology, additional work on phonological awareness and phonics, observation of classrooms in the private school setting during reading and written language classes, and work with numerous case studies to discuss strategies for managing a variety of classroom scenarios.

Design

Woodcock-Johnson III (WJ-III) assessment scores were collected for all students participating in the study. Students were assessed on WJ-III prior to their involvement in the program. After approximately 7 months of participation in the intervention, students were assessed a second time. Students who participated in the intervention a second year were assessed again approximately 12 months after their first posttest assessment. Similarly, students who participated in the intervention a third year were assessed a final time approximately 12 months after their second posttest assessment. While the schedule of assessments for most students followed this pattern, there is some variation in the length of time between the pre- and posttest assessments because children joined the program at different times during the school year. Generally, initial posttest assessments were administered to students after at least 6 months of their involvement in participating classrooms.

The WJ-III assessments were conducted by an independent educational diagnostician with extensive experience in administering a variety of psychological and academic assessments, including WJ-III. Occasionally it was necessary for a teacher of learning disabled or special education students from the home school to administer the assessments to a child. Only teachers with training and the appropriate credentials were permitted to conduct the WJ-III assessments for the purposes of the evaluation.

Measures

The WJ-III Tests of Achievement is a well-known battery of tests designed to measure intellectual abilities and academic achievement. It is used for

educational, clinical, or research purposes; diagnosis; educational programming; guidance; or program evaluation (Woodcock, McGrew, & Mather, 2001). The normative samples (preschool, school aged, college, and adult) were selected to be representative of the U.S. population, based on geographic distribution, community size and type, socioeconomic status, occupational status, gender, and race. For the purposes of this study, only the subset of WJ-III Tests of Achievement measuring different aspects of reading was administered, including the following four tests: Letter-Word Identification, Reading Fluency, Passage Comprehension, and Word Attack. Most of the WJ-III tests show reliability characteristics that meet or exceed basic standards (Woodcock et al., 2001). Taken together, these tests measure various aspects of reading achievement, including reading decoding, reading speed, the ability to comprehend connected discourse while reading, sight vocabulary, phonics, and structural analysis.

Demographic, educational history, and disability status variables were recorded directly from each participating child's school records. School staff were trained to use the study's school records data collection protocol. In addition, attendance data were collected for all participating students each year. The completed forms were mailed directly to RTI by school staff, with appropriate steps taken to ensure the security and confidentiality of student information. Hours of instruction were calculated for each student using the number of minutes per session, number of days per week of sessions, school holidays, and the number of times a student was absent.

Analyses

Preliminary descriptive analyses were used to examine the distributions and missing data patterns of the WJ-III test outcomes. Hours of instructions were used as the measure of time and control items.

The primary analytic model used in this evaluation was piecewise latent growth models (LGMs) (Bollen & Curran, 2006). Descriptive analyses conducted prior to the evaluative models indicated that the change from Year 1 to Year 2 and the change from Year 2 to Year 3 were not equivalent, suggesting that a simple linear model fit through all three assessments was not applicable. Instead, a modified piecewise LGM was estimated with a spline at Year 2 so that individual rates of change could be estimated uniquely for the first and second years of participation in the reading remediation program. Models were estimated in Mplus version 4 (Muthén & Muthén, 1998).

Although Blachman et al. (2004) used LGM in their reading program evaluation, they used a coarse coding in which time was coded simply as an integer reflecting month of assessment (i.e., $t = 0,1,2,3$ for four measurement occasions). Although adequate, this time coding does not reflect individual differences at the student level in amount of instruction

received; thus, each increment of change (the estimated growth coefficient) is reflective of a group-average amount of growth. The data for this study included the hours of instruction each student received each year, enabling models that incorporated this added information. To use these individually varying time variables, all LGMs were estimated in SAS PROC MIXED (SAS Institute, 2005). Models using coarse time coding (0, 1, etc.) were also estimated and effects sizes (Cohen's *d*) and parameters were compared for varying conclusions and implications for the intervention program and schools using the program. Change over time was estimated for all four WJ-III tests. Model intercepts were adjusted for minority status, gender, English as a Second Language status, eligibility for free or reduced lunch, and having ever repeated a grade. All controls were centered to estimate the intercept as the Year 1 overall mean.

RESULTS

Descriptive analyses indicated no problems except for a high percentage (17%) of missing values for IQ. Rather than use multiple imputation (Little & Rubin, 2002) to retain this item, IQ was dropped as a control item that the model intercepts were conditioned on. Correlations with included items indicated substantial overlap of IQ with other control measures with more complete data. Table 2 displays the raw means and standard deviations for each test by year.

Individual Time Coding

LGMs with individual hours of instruction as the time metric indicated that three of the tests showed significant improvement from Year 1 to Year 2 of the intervention. Passage Comprehension increased about 3.13 points, $t(206) = 4.94, p < .001$; Reading Fluency increased approximately 2.2 points, $t(165) = 3.74, p < .001$; and Word Attack scores increased by almost 4.8 points, $t(202) = 6.02, p < .001$. Letter-Word Identification did not show significant improvement, and there was very little change in this test at all

TABLE 2 Raw Means and Standard Deviations

Woodcock-Johnson III test data	Year 1		Year 2		Year 3	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Passage Comprehension	79.99	12.04	83.94	10.09	84.47	10.40
Reading Fluency	81.84	10.86	83.36	10.62	84.79	10.33
Letter-Word Identification	81.43	11.97	82.32	11.70	82.07	7.33
Word Attack	80.86	12.31	86.59	9.52	88.53	9.36

TABLE 3 Model Estimate and Effect Sizes, Individual Time Coding

Woodcock-Johnson test	Label	Estimate	*SE*	*df*	*t*	sig	*d*
Passage Comprehension	Change pre to post 1	3.137	0.635	206	4.94	***	0.651
	Change pre to post 2	2.871	0.860	206	3.34	***	0.453
	Change post 1 to post 2	−0.266	0.850	206	−0.31		−0.044
	Baseline mean	80.322	0.763	141			
	Post 1 mean	83.459	0.777	206			
	Post 2 mean	83.193	0.959	206			
Reading Fluency	Change pre to post 1	2.198	0.588	165	3.74	***	0.559
	Change pre to post 2	4.228	0.888	165	4.76	***	0.695
	Change post 1 to post 2	2.029	0.851	165	2.39	*	0.365
	Baseline mean	80.523	0.806	134			
	Post 1 mean	82.721	0.790	165			
	Post 2 mean	84.751	1.013	165			
Letter-Word Identification	Change pre to post 1	0.127	0.570	205	0.22		0.031
	Change pre to post 2	0.548	0.761	205	0.72		0.100
	Change post 1 to post 2	0.421	0.739	205	0.57		0.080
	Baseline mean	81.211	0.821	141			
	Post 1 mean	81.338	0.844	205			
	Post 2 mean	81.759	0.977	205			
Word Attack	Change pre to post 1	4.789	0.796	202	6.02	***	0.780
	Change pre to post 2	5.861	1.030	202	5.69	***	0.743
	Change post 1 to post 2	1.072	0.960	202	1.12		0.157
	Baseline mean	81.269	0.735	140			
	Post 1 mean	86.058	0.822	202			
	Post 2 mean	87.130	1.040	202			

Note: sig = significance; * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

(.127 points). Significant changes from Year 2 to Year 3 were not as prevalent: only Reading Fluency increased significantly in this time span, by approximately 2.03 points, $t(165) = 2.39$, $p < .05$. Letter-Word Identification and Word Attack increased nonsignificantly; but Passage Comprehension scores showed a slight, nonsignificant decrease from Year 2 to Year 3. When both piecewise change trajectories were combined to estimate the overall change in reading scores from baseline to Year 3, significant increases were found for Passage Comprehension, 2.87 points, $t(206) = 3.34$, $p < .001$; Reading Fluency, 4.23 points, $t(165) = 4.76$, $p < .001$; and Word Attack, 5.86 points, $t(202) = 5.69$, $p < .001$). Effect sizes were sizable, ranging from .45 to .78 for significant effects (medium to large effect size) (Cohen, 1988). Table 3 summarizes model estimates and effect sizes as well as estimated mean scores for each measurement. Model estimated means at each assessment match very closely with the raw observed scores in Table 2 because estimated change was based on only two time points per segment.

Coarse Time Coding

Table 4 summarizes the model estimates and effect sizes using the coarser time coding in which time reflects measurement occasion. The results of

TABLE 4 Model Estimate and Effect Sizes, Coarse Time Coding

Woodcock-Johnson test	Label	Estimate	*SE*	*df*	*t*	sig	*d*
Passage Comprehension	Change pre to post 1	3.943	0.673	206	5.86	***	0.756
	Change pre to post 2	3.228	0.913	206	3.54	***	0.478
	Change post 1 to post 2	−0.715	0.918	206	−0.78		−0.108
	Baseline mean	79.979	0.765	141			
	Post 1 mean	83.923	0.788	206			
	Post 2 mean	83.208	1.001	206			
Reading Fluency	Change pre to post 1	2.586	0.641	165	4.04	***	0.600
	Change pre to post 2	4.450	0.939	165	4.74	***	0.692
	Change post 1 to post 2	1.864	0.897	165	2.08	*	0.319
	Baseline mean	80.342	0.815	134			
	Post 1 mean	82.928	0.798	165			
	Post 2 mean	84.792	1.041	165			
Letter-Word Identification	Change pre to post 1	0.379	0.625	205	0.61		0.085
	Change pre to post 2	0.442	0.789	205	0.56		0.078
	Change post 1 to post 2	0.063	0.706	205	0.09		0.012
	Baseline mean	81.118	0.818	141			
	Post 1 mean	81.498	0.904	205			
	Post 2 mean	81.560	1.024	205			
Word Attack	Change pre to post 1	5.716	0.810	202	7.06	***	0.890
	Change pre to post 2	6.449	1.051	202	6.13	***	0.793
	Change post 1 to post 2	0.733	0.995	202	0.74		0.104
	Baseline mean	80.765	0.732	140			
	Post 1 mean	86.481	0.831	202			
	Post 2 mean	87.214	1.066	202			

Note: sig = significance; * = $p < .05$, ** = $p < .01$, *** = $p < .001$.

these models follow the same pattern as for the individual time-coded models. Passage Comprehension increased significantly from baseline to the first follow-up by approximately 4 points, $t(206) = 5.86$, $p < .001$, and overall from baseline to the Year 3 follow-up, approximately 3.2 points, $t(206) = 3.54$, $p < .001$). Reading Fluency increased from Year 1 to Year 2, approximately 2.6 points, $t(165) = 4.04$, $p < .001$, and again from Year 2 to Year 3, 1.9 points, $t(165) = 2.08$, $p < .05$, for an overall increase of about 4.5 points, $t(165) = 4.74$, $p < .001$). Letter-Word Identification showed little change, as with the previous model. Word Attack increased significantly from Year 1 to Year 2 by approximately 5.7 points, $t(202) = 7.06$, $p < .001$, and overall from Year 1 to Year 3 by almost 6.5 points, $t(202) = 6.13$, $p < .001$.

Although the conclusions reached by the models are the same, the methods have implications for program cost and resource use as related to per-hour increases. The coarse time measure ignores individual-level "dose" or instruction time, and essentially uses the average across all participants as the time amount. The ramification for policy and resource allocation is that this strategy can be influenced by deviations from normality assumptions, such as when there are outliers in hours received or markedly above- or below-average performers. Table 5 compares the estimated change per hour

TABLE 5 Comparison of Estimated Change per Hour

Woodcock-Johnson III test	Points/hour, coarse	Points/hour, individual	Ratio
Passage Comprehension	0.070	.056	1.25
Reading Fluency	0.046	.039	1.18
Letter-Word Identification	0.007	.002	3.50
Word Attack	0.102	.086	1.19

for the coarse and individual hours of instruction models. Overall change in the coarse model is divided by the average number of instructional hours (56.01 from Year 1 to 2). The time parameter estimate pertaining to Year 1 to 2 change in the individual model represents the observed change per hour in this model. In addition to greater change per hour, the coarse coding resulted in somewhat larger effect sizes.

To illustrate the impact of these differing model results, consider the total gains expected, on average, if each child received the recommended number of instructional hours in a year. The program recommends 100 instructional hours per year per child, which would yield an expected gain of 7 points on the WJ-III Passage Comprehension test using the coarse time model. However, the model that incorporates individually varying exposure to the program yields a predicted gain of 5.6 points. This difference is most glaring for the Letter-Word Identification test, for which the coarse model predicts 0.7 points increase for 100 hours, whereas the individual time model predicts a 0.2 points increase. Such a difference greatly affects the converse consideration: educators would allocate too few resources (e.g., time) to achieve a desired gain if guided by the results of the coarse time model.

DISCUSSION

The intervention evaluated in this study was an adaptation of a remediation methodology created specifically as a public school program in which struggling readers are taken out of their regular classrooms daily for short periods to receive specialized instruction. The program addressed the five essential components of a successful reading program, as put forth in the National Reading Panel Report of 2000 (NICHD, 2000). However, the teaching methods are not those typically used in traditional classroom settings.

The intervention involved a multisensory structured language approach. According to the International Dyslexia Association (2001), this type of approach typically teaches phonology and phonological awareness, sound-symbol association, syllable instruction, morphology, syntax, and semantics. It has been found to be effective with learning disabled students. There also is an extensive body of research that reading interventions

emphasizing phonology and phonological awareness are effective in improving the reading skills of struggling readers (Blachman, 2000; Lyon, 1998, 2002; NICHD, 2000; Torgesen et al., 2006).

One goal of this study is to examine the relationship between an important measure of implementation fidelity—dosage of or exposure to the intervention—and the primary study outcome: the reading achievement of students at risk for failure in reading in a public school setting. Including measures of implementation fidelity in evaluations of school-based interventions and programs is critical to understanding how the degree of program implementation can affect the achievement of desired goals—in this instance, improved reading achievement. Studying implementation fidelity will also help educators and program developers understand how to improve implementation over time as well as in alternative or broader contexts.

In this study, intervention dosage or exposure was measured for each participating student. Using this measure of implementation fidelity, it was possible to examine the variance between recommended and received hours of instruction. The study findings indicate that, on average, students who received greater intervention exposure showed higher levels of improvement in reading achievement. However, the findings also show that despite the fact that, on average, students received less than the recommended dosage of the intervention, significant improvement in reading achievement occurred. This is an important finding for educators because it shows that the intervention had a positive impact, despite the fact that the intervention was not fully implemented in terms of dosage. In addition, the findings suggest that further improvement in achievement may take place if students are provided greater intervention exposure.

As part of examining the relationship between intervention dosage and students' reading achievement levels, in analytic modeling for change over time (e.g., growth models) we also tested the impact of using individual-level versus aggregate-level data on the extent of intervention exposure. More specifically, the analysis presented compared two different models for examining the relationship between intervention exposure and change in reading achievement. One model used simple time coding (i.e., average hours of instruction for all students in a given grade level) as the measure of intervention exposure; the second model used individual time scores (i.e., actual number of hours of the intervention each student received).

The findings show that the choice of model has important implications for estimates of the impact of the intervention on achievement. More specifically, when individual levels of exposure to or dosage of the intervention are ignored, instead opting for an analytic method that relies on the average level of exposure across all subjects, the effect size may be influenced by outliers (i.e., above- or below-average performers). In this case, estimates for the number of hours of exposure or treatment that the average student needs to show significant gains may be inaccurate (e.g., too high or too low). In turn,

information used by administrators to determine cost and resource allocation relative to anticipated results may be inaccurate. In addition to minimizing these types of interpretive error, analytic modeling that takes into account individual-level measures of dosage or exposure offers practitioners a strategy for more closely monitoring individual students' responsiveness to the intervention.

Evaluations of intervention programs are often limited by the quantity of data obtained. Longitudinal designs are especially pressed, as there is demand not only for adequate sample size but also for multiple assessments so as to accurately estimate change over time. Unfortunately, in many cases, even multiple assessments are a limited proxy for the true underlying time dosage (Singer & Willett, 2003). In this evaluation, the waves (1, 2, and 3) do not indicate the actual underlying chronological element, but rather the receipt of instruction. Conclusions based on measurement wave alone essentially limit the evaluation to estimates of group-averaged effects, ignoring how many or few hours of the intervention a student actually received. By using hours of instruction as the indicator of time in the models, a more accurate estimate of the change per hour of instruction is obtained. Although the conclusions from both models are comparable in terms of significance tests, the estimated change per hour has important implications for what happens after the program evaluation is completed and further use of the intervention is deemed appropriate.

Limitations

When considering the results of this study, it is important to note several limitations. First, the evaluation was conducted on a relatively small sample, with no more than approximately 140 students participating at any of the three assessments. Typically, small samples are associated with lack of power, but this seems to be a consideration for only the Letter-Word Identification test. The other three reading achievement tests showed significant change over time, with considerable effect sizes. A related concern is that the small sample limits the generalizability of the study findings. Smaller samples pose a greater risk of idiosyncratic effects based on sample composition.

Perhaps the greatest concern regarding intervention effectiveness is that program students were not compared with a control group of similar students who did not receive the intervention. However, two design features somewhat mitigate against this concern. First, the outcomes analyzed are standard scores for each year, such that each student, had he or she achieved the desired year's worth of reading ability increase, would show a flat longitudinal trajectory. This is because each year a student who had acquired the yearly amount of reading would receive the same standard score as the year before. This, combined with the repeated measurement of study participants, results

in a situation where the assumed, or comparative, growth is zero (which is the expected growth for a random sample of control participants) and any deviations in slope over time reflect either accelerated or decelerated rates of reading ability acquisition. The significant positive slopes found in this study indicate that the participants acquired reading skills at a pace that was accelerated relative to an expected yearly amount (i.e., their skill increased at a rate that exceeded the expected zero slope standard score change).

REFERENCES

Blachman, B. A. (2000). Phonological awareness. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 483–502). Mahwah, NJ: Erbaum.

Blachman, B. A., Schatschneider, C., Fletcher, J. M., Francis, D. J., Clonan, S. M., Shaywitz, B. A., et al. (2004). Effects of intensive reading remediation for second and third graders and a 1-year follow-up. *Journal of Educational Psychology, 96*, 444–461.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: John Wiley & Sons.

Cohen, J. (1988). *Statistical power for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Dusenbury, L., Brannigan, R., Mathea, F., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research, 18*, 237–256.

Erlbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-on-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology, 92*, 605–619.

International Dyslexia Association. (2001). *Orton-Gillingham-based and/or multisensory structured language approaches* [Fact Sheet #968]. Retrieved from http://www.wrightslaw.com/info/read.msl.ida.pdf#search=%22ida%20orton-gillingham-based%22

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley.

Lyon, G. R. (1998).*The NICHD research program in reading development, reading disorders and reading instruction: A summary of research findings*. New York: National Center for Learning Disabilities, Inc.

Lyon, G. R. (2002). *Testimonies to Congress: 1997–2002*. Covington, LA: Center for the Development of Learning.

Melde, C., Esbensen, F., & Tusinski, K. (2006). Addressing program fidelity using onsite observations and program provider description of program delivery. *Evaluation Review, 30*(6), 714–740.

Mokrue, K., Elias, M., & Bry, B. (2005). Dosage effect and the efficacy of a video-based teamwork-building series with urban elementary school children. *Journal of Applied School Psychology, 21*(1), 67–97.

Moody, S. W., Vaughn, S., Hughes, M. T., & Fischer, M. (2000). Reading instruction in the resource room: Set up for failure. *Exceptional Children*, *66*, 305–316.

Muthén, L. K., & Muthén, B. O. (1998–2006). *Mplus user's guide* (4th ed.). Los Angeles: Muthén & Muthén.

National Center for Education Statistics. (2007). *Common core of data, public school data, 2004–2005 school year*. Retrieved April 10, 2007, from http://nces.ed.gov/ccd/

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching children to read: Reports of the subgroups* (00-4754). Washington, DC: U.S. Government Printing Office.

Ruiz-Primo, M. (2006). *A multi-method and multi-mource approach for studying fidelity of implementation*. CSE Report 677. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE).

SAS Institute. (2005). SAS (Version 9.1) [Computer software]. Cary, NC: SAS Institute.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Methods for studying change and event occurrence*. New York: Oxford University Press.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Torgesen, J., Stancavage, F., Myers, D., Schirm, A., Stuart, E., Vartivarian, S., et al. (2006). *Closing the reading gap: First year findings from a randomized trial of four reading interventions for striving readers*. Washington, DC: Corporation for the Advancement of Policy Evaluation.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., et al. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology*, *91*, 579–593.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R. S., et al. (1996). Cognitive profiles of difficult to remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology*, *88*, 607–638.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of achievement*. Itasca, IL: Riverside Publishing.