Journal of Educational Psychology
2004, Vol. 96, No. 3, 444–461

0022-0663/04/$12.00 DOI: 10.1037/0022-0663.96.3.444

Effects of Intensive Reading Remediation for Second and Third Graders and a 1-Year Follow-Up

Benita A. Blachman
Syracuse University

Christopher Schatschneider
Florida State University

Jack M. Fletcher
University of Texas Health Science Center at Houston

David J. Francis
University of Houston

Sheila M. Clonan
Syracuse University

Bennett A. Shaywitz and Sally E. Shaywitz
Yale University Medical School

Second- and 3rd-grade children with poor word-level skills were randomly assigned to 8 months of explicit instruction emphasizing the phonologic and orthographic connections in words and text-based reading or to remedial reading programs provided by the schools. At posttest, treatment children showed significantly greater gains than control children in real word and nonword reading, reading rate, passage reading, and spelling, and largely maintained gains at a 1-year follow-up. Growth curve analyses indicated significant differences in growth rate during the treatment year, but not during the follow-up year. Results indicate that research-based practices can significantly improve reading and spelling outcomes for children in remedial programs.

Considerable attention has been paid recently to the benefits of early intervention for children who are struggling to learn to read (Snow, Burns, & Griffin, 1998). The evidence is compelling that early reading programs that emphasize the connections between the phonological structure of spoken words and the alphabet can help close the gap between children who struggle to learn to read and those who learn to read easily (Blachman, 2000; Ehri, Nunes, Stahl, & Willows, 2001; Liberman & Liberman, 1990; Vellutino, 1991). Although knowledge regarding the reduction of reading failure has grown (Adams, 1990; Stanovich, 1986; Torgesen, Wagner, Rashotte, Rose, et al., 1999; Vellutino et al., 1996), more empirical research has been devoted to preventing failure in at-risk children, especially those in kindergarten and first grade, than to investigating the effectiveness of explicit and intensive instruction for children selected after they have already exhibited reading failure (Elbaum, Vaughn, Hughes, & Moody, 2000; National Reading Panel [NRP], 2000). In addition, the evidence suggests that prevention studies have generally been more successful, as evidenced by larger effect sizes on reading and spelling measures, than studies of children who have already experienced reading failure (Lyon et al., 2001; Torgesen, 2000).

The purpose of this study was to evaluate the effectiveness of an intensive reading intervention for second- and third-grade children with reading disabilities who were selected on the basis of poor word-level skills. Our goals included developing an intervention that would target both accuracy and fluency and evaluating our model to determine whether the growth trajectory of these second- and third-grade children would be altered by an intensive intervention. To address the limited word recognition skills of these children, our intervention incorporated both explicit skill-based instruction and frequent opportunities for text-based reading. The need for young children to gain accurate and fluent word-level skills has been reinforced by many researchers (Adams, 1990; Ehri, 1991; Perfetti, 1985; Pressley, 2002; Share & Stanovich, 1995; Vellutino, Scanlon, & Tanzman, 1994; Williams, 1994), as well as by two influential consensus reading panels, the first convened by the National Research Council (Snow et al., 1998) and the second commissioned by Congress (NRP, 2000). Snow et al. (1998) concluded:

> The first obstacle, which arises at the outset of reading acquisition, is difficulty understanding and using the alphabetic principle—the idea that written spellings systematically represent spoken words. It is hard

Benita A. Blachman, Reading and Language Arts Center and Department of Psychology, Syracuse University; Christopher Schatschneider, Department of Psychology, Florida State University; Jack M. Fletcher, Department of Pediatrics, University of Texas Health Science Center at Houston; David J. Francis, Department of Psychology, University of Houston; Sheila M. Clonan, Reading and Language Arts Center, Syracuse University; Bennett A. Shaywitz, Department of Pediatrics and Department of Neurology, Yale University Medical School; Sally E. Shaywitz, Department of Pediatrics, Yale University Medical School.

This research was one of several projects supported by Center Grant P50HD25802 awarded to Yale University School of Medicine by the National Institute of Child Health and Human Development. Bennett A. Shaywitz and Sally E. Shaywitz were codirectors of the Center grant. The specific study reported in this article was funded by a subcontract of Center Grant P50HD25802 directed by Benita A. Blachman at Syracuse University. We wish to thank the schools, children, parents, teachers, tutors, and testers who participated in this research. We also express our sincere appreciation to Maria Murray, Mindy Scirri, Rachel Karchmer, and Darlene Tangel for their contributions to this project.

Correspondence concerning this article should be addressed to Benita A. Blachman, Syracuse University, 200 Huntington Hall, Syracuse, NY 13244–2340. E-mail: blachman@syr.edu

> to comprehend connected text if word recognition is inaccurate or laborious. (p. 4)

The need to change the growth trajectory, specifically of second- and third-grade children with reading disabilities, is important for at least two reasons. First, epidemiological data indicate that if children's reading skills haven't improved by the end of third grade, these children will have considerable difficulty overcoming their slow and unsuccessful start in reading. Specifically, S. E. Shaywitz and Shaywitz (l996) found that 74% of children who were still poor readers at the end of third grade were likely to be poor readers at the end of ninth grade. Negative economic and emotional consequences often follow poor readers into adulthood (Lyon, 2001). Researchers have suggested that "what seems essential is to insure that children learn to decode in first grade. If decoding skill does not arrive then, it may be very hard to change the direction that reading achievement takes" (Gough & Juel, 1991, p. 55). Data included in the NRP (2000) report indirectly support the Gough and Juel suggestion. The NRP found significantly smaller effect sizes for both reading and spelling outcomes in studies of "older" disabled readers (i.e., Grades 2–6) who were learning phonics compared with studies of kindergarten and first-grade children learning phonics (although see Rashotte, MacPhee, & Torgesen, 2001). As outlined eloquently by Stanovich (1986) more than 15 years ago, children who are not successful learning to decode, or read words, in first grade, read less often and fail to develop the fluency that comes with reading practice. When reading words is slow and laborious, cognitive resources that might otherwise be devoted to meaning must be used in the service of sounding out words. In addition, these same children are at a disadvantage in terms of the vocabulary acquisition and world knowledge accumulated by those who learn to read easily and who read often (Cunningham & Stanovich, 1997).

A second reason for focusing on second and third graders is that the instructional focus in reading changes rather drastically in fourth grade, putting children who are still poor readers at a considerable disadvantage. As articulated by Snow et al. (1998), "In first grade, the challenge for children is to learn how to read. In fourth grade and up, it is taken for granted that they are capable . . . of reading to learn" (p. 207). This change in instructional focus makes it less likely that fourth-grade children will routinely be exposed to explicit instruction in reading. For children in second and third grades who have not met the first-grade "learning to read" challenge, explicit and systematic reading instruction that emphasizes the phonologic and orthographic connections in words may be especially important. There is little evidence, however, that such instruction is routinely provided. Many children, despite limited reading skills, are not getting the help they need at the time they could most profit from such help because of the peculiarities of the discrepancy formula used to label children learning disabled. Children who have not yet met the required discrepancy between IQ and achievement must continue to "wait-to-fail," falling even further behind before getting assistance (Lyon & Fletcher, 2001; Stuebing et al., 2002). At the same time, many children who are receiving remedial reading support in schools are reportedly making little or no progress, or, as Torgesen et al. (2001) put it, the programs are merely "stabilizing their degree of reading failure" (p. 34). Researchers have long documented the lack of effectiveness of two common sources of remedial reading instruction provided to public school children: Chapter 1 reading programs and reading instruction provided in the resource room for children with disabilities (Kennedy, Birman, & Demaline, 1986; Moody, Vaughn, Hughes, & Fischer, 2000; Puma et al., 1997; Snow et al., 1998), reporting that any gains that are made are lost when children leave the program (Birman et al., 1987).

In this study our goals were to (a) evaluate an intervention for second- and third-grade students with poor word-level skills; (b) monitor their progress for 1 year after the intervention ended to investigate whether gains were maintained; and (c) determine which areas of reading and spelling, if any, demonstrated long-term gains. Although all children began the study with word-level skills below the 25th percentile on a standardized test, skill levels actually ranged from the 1st to the 23rd percentile. Given this range, we were also interested in exploring the effectiveness of the intervention for children who began the study with more word reading skill and those who began the study with less skill in word reading.

Method

Participants

Children were selected from 11 schools in four school districts in upstate New York. There were two cohorts of children. Those in the first cohort were recruited in the spring of 1997 from 6 participating schools, and those in the second cohort were recruited in the spring of 1998 from 11 participating schools. Cohort 2 schools included the 6 schools from which Cohort 1 children were drawn and 5 new schools from the same districts. The districts represented a range of socioeconomic status levels from poor, urban schools to middle-class, suburban schools. Data from the two cohorts have been aggregated in the remainder of the article.

It should be noted that the treatment children selected for this study were also part of a study to investigate the influence of an intensive reading intervention on patterns of brain activation in children with reading disabilities (B. A. Shaywitz et al., 2004). Because of the neuroimaging component, selection criteria were somewhat more constrained than is typical in an intervention study. For example, to be included in the study, children had to be right-handed and could not have medical appliances (e.g., pacemaker) or other devices (e.g., braces) that might preclude their having a magnetic resonance imaging (MRI) scan.

The children were selected during a two-stage process. During the first stage, in the spring of the school year, first- and second-grade teachers in participating schools were asked to identify children they considered to be among the lowest 20% of readers in their classrooms, excluding children who were left-handed; who had hearing loss, severe articulation problems, severe emotional disturbance, autism, mental retardation, or neurological problems (e.g., epilepsy); or who were learning English as a second language. To select the children during this first stage, schools could use procedures already in place for identifying children in need of special services (e.g., Chapter 1, Reading Recovery, resource room instruction for those labeled learning disabled) or a scale developed by Vellutino et al. (1996) that asked teachers to rate children in reading from 1 (*low*) to 5 (*high*) using the following criteria: 1 = *child is having a great deal of difficulty in reading*, 2 = *child is having some degree of difficulty in reading*, 3 = *child is making normal progress*, 4 = *child is progressing somewhat more quickly than normal*, and 5 = *child is progressing extremely well*. Permission to screen the school-identified children for participation in this stage of the study was obtained by having the schools send letters to parents, thus protecting the confidentiality of the families. Parents willing to have their children screened on a reading and IQ test returned their permission letters directly to Benita A. Blachman. These parents were contacted by the research team, and the initial screening was scheduled.

Letters were sent to the families of 723 children, or 21% of the first and second graders in the participating schools. Permission for screening was received from 306 parents, or 42%[1] of the school-identified children. Of the 306 children for whom permission was obtained, we screened 295 children. Prior to screening, 11 children were eliminated for the following reasons: 7 were left-handed, 2 were learning English as a second language, 1 moved before testing, and 1 was repeatedly absent on days scheduled for screening.

Children were first given a reading test. All children who obtained a standard score below 90 on either the Word Identification or the Word Attack subtest of the Woodcock Reading Mastery Tests—Revised (WRMT–R; Woodcock, 1987) and who also obtained a standard score below 90 on the Basic Skills Cluster (a composite of the Word Identification and Word Attack subtests) of the WRMT–R were given the Wechsler Intelligence Scale for Children—Third Edition (WISC–III; Wechsler, 1991). Children who met the reading criteria and who also had a Verbal IQ score of at least 80 were considered eligible for the study. Although the teachers originally nominated children they felt were in need of help in reading, our criteria turned out to be more stringent than those used by the school. Of the 295 children screened, 156 (52%) were eliminated because they had a Basic Skills Cluster on the Woodcock of 90 or above and, consequently, were not considered reading impaired according to our criteria. An additional 25 children (8%) had WISC–III scores below the cut-off, 5 were eliminated because they were outside the upper age limit for inclusion in the study (i.e., 8 years 11 months at the time of screening), 3 were determined to be left-handed, 2 children exhibited behavior during testing that was suggestive of a more pervasive developmental disorder, 2 had severe neurological disorders, 1 was learning English as a second language, and 1 moved before completing the screening.

Once the screening was completed and children were eliminated for the reasons described above, 100 children (65 male, 35 female) remained eligible for the study. Eleven boys (17% of the eligible boys) were randomly eliminated from the eligible pool. This decision was made in an attempt to bring the proportion of girls closer to the proportion of boys for the neuroimaging component of the study. The remaining 89 students (54 male, 35 female) were randomly assigned within schools, grade, and gender to treatment and control groups (e.g., if there were 2 second-grade boys at a given school, 1 was randomly assigned to the treatment group and 1 was randomly assigned to the control group).

We then began the second stage of the selection process, asking parents for permission for their child to participate in the group to which the child had been randomly assigned. Although it might be more common to get permission from families for their child to participate as either a treatment or control child and then randomly assign, this procedure was not considered appropriate in this study for the following reason. The treatment children in this study were participants in a larger neuroimaging study that required three out-of-state trips to New Haven, Connecticut (for up to 3 days each) for neuroimaging and other assessments. Because of these complicated logistics, we found it necessary to have extensive conferences with each potential treatment family. In addition, most parents of children assigned to the treatment group needed time to confer with extended family, including the identified child and employers (regarding time off), before making a decision regarding participation. We were not comfortable asking all 89 families to go through this extensive process regarding participation (including introducing the neuroimaging concept to a 7- or 8-year-old child), only to tell half of the families and children that, although they agreed to the neuroimaging and out-of-state trips, none of this would happen. Consequently, in this second stage of the selection process, we randomly assigned children to conditions and then contacted parents to discuss their interest in having their child participate in the next phase of the study and to get informed consent.

We initially overassigned children to the treatment group because we thought fewer parents would agree to the out-of-state travel required for the neuroimaging component of the study. However, among the 48 children originally assigned to the treatment group, only one family declined because of this component of the study. In addition, three parents failed to show up for either of two scheduled conferences and were not contacted further, three parents indicated that they would be moving prior to the start of the new school year, 1 child was eliminated because of a serious heart condition that made him ineligible for the neuroimaging component of the study, and 1 child was eliminated because of second language issues. Two children were excluded from the treatment group because of extensive metal dental work, making them ineligible for the MRI. Both of these latter children were reassigned to the control group. This resulted in a treatment group of 37 children, including 22 boys and 15 girls. Of the 37 children in the group, 28 of these children were White, 6 were African American, 1 was Hispanic, and 2 were identified by the parents as Other.

Forty-one children were randomly assigned to the control group. Repeated attempts to reach one family were unsuccessful, three parents indicated that they would be moving prior to the start of the new school year, and six families with whom we discussed the study declined to participate. Several parents who declined to participate expressed concern that their child would not be receiving tutoring when others in the study would be tutored. As we indicated previously, 2 children, described above, were reassigned from the treatment to the control group, and 1 of those children agreed to participate. This resulted in a control group of 32 children, including 20 boys and 12 girls. Of these 32 children, 27 were White, 4 were African American, and 1 was identified by the parents as Other.

No significant differences were found between the treatment and control groups on age, $t(67) = -1.00$, $p = .32$; sex, $\chi^2(1, N = 69) = 0.07$, $p = .80$; race (Fisher's exact test, $N = 69$, $p = .94$); mother's educational level, $t(67) = 0.34$, $p = .73$, or any of the initial screening measures used to determine eligibility, as seen in Table 1, including Verbal IQ, Performance IQ, Full Scale IQ, Woodcock Basic Skills Cluster, Woodcock Word Identification subtest, or the Woodcock Word Attack subtest. There was no attrition during this study, and all analyses are based on 37 children in the treatment group and 32 children in the control group. In addition, the groups did not differ on the number of school absences during the treatment year, $t(67) = 0.24$, $p = .81$, or the follow-up year, $t(67) = 1.01$, $p = .30$.

Measures

When the children were beginning Grade 2 or 3, both groups of children were given two batteries of tests. One battery consisted of standardized (norm-referenced) tests, including measures of reading, spelling, and math that the children completed three times—as pretests before tutoring began; as posttests at the end of tutoring; and again, 1 year after the posttest, at follow-up. The math measures were included as a control to determine whether our intervention actually targeted reading, as planned. We hypothesized that a finding of posttest group differences in reading, but no differences in math, would provide an additional source of evidence that differences in reading were, in fact, due to the intervention. A second

[1] As noted, only 42% of parents contacted agreed to have their children screened. Although some self-selection bias might have been operating (e.g., the most concerned parents might have been more likely to return the letter), this does not bias the treatment effect because we randomized children to treatment and control conditions within each school. We found no relationship in the sample between return rate and socioeconomic status. (The highest return rate from a single school—59%—was from the school with the highest percentage—61%—of children receiving free or reduced-price lunch.) In addition, there is no evidence that this return rate reduces the generalizability of the sample, because 8% of the eligible pool (i.e., 42% of the 20% of eligible children) responded, a percentage not inconsistent with population percentages of disabled readers (Snow et al., 1998). Our initial parent letter may have influenced our response rate. We asked for permission to screen but did not indicate that screening might lead to tutoring. Including the potential for tutoring in our initial letter might have increased our response rate.

Table 1
Pretest, Posttest, Follow-Up Means and Adjusted Means for Treatment and Control Groups

	Treatment[a]			Control[a]					
Measure	*M*	*SD*	*M*[b]	*M*	*SD*	*M*[b]	*F*	*p*	*d*[c]
Pretest only									
Age at entry (years)	7.95	0.49		7.82	0.57		1.00	0.3219	
WISC–III Verbal IQ	94.81	10.07		92.88	9.17		0.69	0.4095	
WISC–III Performance IQ	99.16	13.40		98.47	14.16		0.04	0.8352	
WISC–III Full Scale IQ	96.38	10.58		95.13	11.41		0.22	0.6375	
Pretest									
WRMT Basic Skills Cluster	81.89	6.99		82.38	6.22		0.09	0.7642	
WRMT Word ID	82.81	6.83		83.97	6.45		0.56	0.4740	
WRMT Word Attack	83.30	8.69		81.97	7.32		0.46	0.4984	
WRAT Spelling	82.08	6.52		81.44	6.95		0.16	0.6930	
GORT Quotient	73.08	8.08		74.03	7.01		0.27	0.6063	
GORT Accuracy	77.43	6.30		78.91	6.81		0.86	0.3541	
GORT Rate	76.62	5.66		78.13	5.79		1.19	0.2800	
GORT Comprehension	78.24	11.56		78.75	10.00		0.04	0.8474	
WJ–R Calculations	89.89	12.71		88.63	14.87		0.14	0.7039	
WJ–R Applied Problems	99.22	15.25		95.94	11.81		0.98	0.3273	
Posttest									
WRMT Basic Skills Cluster	88.32	12.60	88.62	78.88	9.59	78.53	33.64	0.0001	1.69
WRMT Word ID	88.65	12.09	89.33	80.38	9.83	79.59	30.58	0.0001	1.31
WRMT Word Attack	90.16	14.07	89.67	80.03	9.46	80.61	13.10	0.0006	0.89
WRAT Spelling	92.22	8.48	92.03	82.81	9.57	83.03	21.62	0.0001	1.13
GORT Quotient	84.92	11.28	85.18	77.78	10.25	77.48	10.24	0.0021	0.78
GORT Accuracy	87.57	13.00	88.07	80.63	10.83	80.04	8.82	0.0041	0.72
GORT Rate	84.05	8.88	84.60	77.97	8.41	77.34	15.84	0.0002	0.96
GORT Comprehension	89.73	11.96	89.80	83.75	11.29	83.67	5.06	0.0275	0.55
WJ–R Calculations	94.76	16.51	94.40	99.00	18.28	99.41	1.82	0.1803	−0.33
WJ–R Applied Problems	102.92	12.40	101.96	104.03	11.81	105.14	2.34	0.1316	−0.37
Follow-up									
WRMT Basic Skills Cluster	87.57	11.68	87.82	79.56	11.29	79.27	16.24	0.0001	0.97
WRMT Word ID	87.30	10.94	87.87	79.63	11.03	78.97	18.75	0.0001	1.05
WRMT Word Attack	89.46	12.06	88.97	82.34	13.08	82.91	5.34	0.0243	0.56
WRAT Spelling	90.59	9.95	90.35	83.56	9.42	83.84	11.02	0.0015	0.81
GORT Quotient	84.84	11.64	85.22	79.94	12.23	79.50	5.52	0.0218	0.57
GORT Accuracy	81.89	12.93	82.37	78.91	14.59	78.36	1.56	0.2151	0.30
GORT Rate	82.43	12.51	83.10	75.78	9.76	75.01	11.09	0.0014	0.81
GORT Comprehension	92.57	11.64	92.67	90.00	12.89	89.88	1.02	0.3163	0.24
WJ–R Calculations	98.00	14.69	97.79	92.88	11.77	93.12	2.37	0.1289	0.38
WJ–R Applied Problems	104.54	9.34	103.93	103.56	11.48	104.26	0.03	0.8760	−0.04

Note. WISC–III = Wechsler Intelligence Scale for Children—Third Edition; WRMT = Woodcock Reading Mastery Tests—Revised; WRAT = Wide Range Achievement Test 3; GORT = Gray Oral Reading Tests—Third Edition; WJ–R = Woodcock-Johnson Psycho-Educational Battery—Revised.
[a] Treatment group $n = 37$; control group $n = 32$. [b] Means adjusted by pretest. [c] d = Cohen's measure of effect size.

battery, referred to as our skills battery, was repeated every 10 weeks (four times during the treatment year and four times during the follow-up year, in September, December, March, and June) and included nonstandardized (nonnormed) measures of phonological processing, word reading accuracy, Word Reading Efficiency (a timed reading measure), and spelling. The order of administration of tests in both batteries was fixed across all administrations. The battery of standardized tests took approximately 1 hr to administer, and the battery of nonnormed tests took approximately 45 min. All testers were extensively trained in the administration and scoring of both batteries, retrained before each wave of testing, and blind to the condition of the children.

Pretest, Posttest, Follow-Up Battery (Each Administered Three Times)

Woodcock Reading Mastery Tests—Revised (WRMT–R), Word Identification subtest and Word Attack subtest (Form G; Woodcock, 1987). The Word Identification subtest measures word recognition by having children read words from a graded word list. For consistency across subjects, testing began with Item 1 for all children, despite recommendations in the manual for the use of different starting points for children in different grades. The median split-half reliability for the Word Identification subtest (Form G) is .97 (Woodcock, 1987). The Word Attack subtest measures word attack skills by having children read decodable nonwords. The median split-half reliability for the Word Attack subtest (Form G) is .87 (Woodcock, 1987).

Gray Oral Reading Tests—Third Edition (GORT–3, Form A; Wiederholt & Bryant, 1992). The GORT–3 measures reading accuracy, rate, and comprehension through timed reading of up to 13 graded passages. Testing yields individual subtest scores as well as an overall oral reading quotient. For consistency, all students began with the first passage on all administrations of the test, despite recommendations in the manual for the use of different starting points for children in different grades. Median internal consistency across all ages ranges from .87 for the comprehension subtest to .97 for the oral reading quotient (Wiederholt & Bryant, 1992).

Wide Range Achievement Test 3, Spelling (WRAT3; Wilkinson, 1993). The WRAT3 Level 1 spelling subtest was individually administered to assess the student's ability to write single words from dictation. Median internal consistency across all ages for the spelling subtest (tan form) is .89 (Wilkinson, 1993).

Woodcock-Johnson Psycho-Educational Battery—Revised (WJ–R), Tests of Achievement, Calculation subtest, and Applied Problems subtest (Woodcock & Johnson, 1989). The Calculation subtest measures the child's ability to perform standard calculations (e.g., addition, subtraction) using pencil and paper to provide answers to problems presented in a traditional format. Internal consistency for the calculation subtest is .93 (Woodcock & Mather, 1989). The Applied Problems subtest measures the child's ability to solve word problems read by the examiner. Internal consistency for the applied problems subtest is .91 (Woodcock & Mather, 1989).

Skills Battery (Each Administered Eight Times)

Blending Phonemes—Words (prepublication version of the Comprehensive Test of Phonological Processes [CTOPP]; Wagner, Torgesen, & Rashotte, 1999). In this task, the child listened to 20 words segmented by phonemes (e.g., /m/ /a/ /n/) and was asked to blend the sounds together to produce a word.[2] Testing began with four training trials that provided modeling and corrective feedback. Testing continued until four out of seven items were missed. The internal consistency estimate measured at the beginning of the treatment year (i.e., the first wave of data collection in September) was .80.

Blending Phonemes—Nonwords (prepublication version of the CTOPP). In this task, the child listened to 15 segmented nonsense words (e.g., /f/ /e/ /t/). The child was asked to put the sounds together to make a nonsense word. Testing began with four practice items and was discontinued when the child missed four out of seven items. The internal consistency estimate measured at the beginning of the treatment year was .82.

Segmenting Phonemes (prepublication version of the CTOPP). The 20 items on this subtest were real words (e.g., *man*) spoken normally by the examiner. The child was asked to segment each word into individual phonemes. The test items were again preceded by four training trials, and testing continued until four out of seven items were missed. The internal consistency estimate for this subtest measured at the beginning of the treatment year was .86.

Phoneme Elision (prepublication version of the CTOPP). This orally presented task required the child to say the word that would be left when a specified phoneme was deleted (e.g., *meat* without the /t/). Twenty items were pronounced by the examiner, and the child was asked to delete initial, final, and medial sounds. The task was preceded by four training trials and discontinued when four out of seven items were missed. The internal consistency estimate for this subtest measured at the beginning of the treatment year was .88.

The four measures of phonological awareness (Blending Words, Blending Nonwords, Segmenting Phonemes, and Phoneme Elision) were summed into a single composite measure. The internal consistency estimate for this composite phonological awareness measure was .84.

Nonword Repetition (prepublication version of the CTOPP). In this task, the child was asked to repeat 25 nonsense words presented orally by the examiner. Three training trials were presented before the test items began. The test was discontinued when the child failed to repeat correctly five consecutive items. The internal consistency estimate for this subtest measured at the beginning of the treatment year was .76.

Rapid Naming of Letters. In this measure, the child was asked to name rapidly five lowercase letters (*o, a, s, d, p*) originally used in studies by Denckla and Rudel (1976). The letters were displayed in five horizontal rows of 10 items per row, and the child was given 60 seconds to name as many letters as possible. The task was preceded by a demonstration trial in which the child was asked to identify each of the five individually presented letters and given the correct name for the letter if an error was made. The test–retest reliability was .78, as measured from the first data collection of the treatment year (September) to the second data collection of the treatment year (December).

Word Reading (modified by B. Foorman and C. Schatschneider from a task used in Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998). The child was asked to read 41 words graded in difficulty. The words were presented individually on flashcards, and testing was discontinued when the child missed 7 consecutive words. The internal consistency reliability of this 41-word test, measured during the first test administration, was .95.

Word Reading Efficiency (prepublication version of the Test of Word Reading Efficiency; Torgesen, Wagner, & Rashotte, 1999). On this task, the child was asked to read as quickly as possible a graded list of 104 real words (Form A), skipping over any unknown words. After 45 seconds, the examiner instructed the child to stop and a second timed trial with a new, but comparable, list of 104 words (Form B) was administered. The timed trials were preceded by a practice trial during which the child was asked to read 8 words as quickly as possible. Speed (number of correctly read words per second) was computed for each version of the test, and the two speed measures were averaged. Test–retest reliability of this average, as measured from the first data collection of the treatment year (September) to the second data collection of the treatment year (December), was .95.

Spelling Dictation (modified by B. Foorman and C. Schatschneider from a measure used in Foorman, Francis, Fletcher, & Lynn, 1996). The child was asked to spell a list of 50 words, 25 of which also appear on the Word Reading measure described above. Testing was discontinued after 7 consecutive misspelled words. The internal consistency reliability estimate obtained from the first wave of data collection was .88.

Procedures

Treatment Year

Treatment children received 50 min of one-to-one tutoring, 5 days per week, between the pretests administered at the end of September and the posttests administered at the beginning of June. These tutoring sessions

[2] During the treatment year only, the first cohort of 21 children (12 treatment; 9 control) were presented 15 items (not 20) on three of the phonological awareness measures, specifically, Blending Phonemes—Words, Segmenting Phonemes, and Phoneme Elision. Because of concerns about ceiling effects at the end of the treatment year, 5 additional items were added to each of these three measures, increasing the number of items on each measure to 20 and increasing the difficulty of each measure. The new items for each measure were selected from a prepublication version of the CTOPP (Wagner et al., 1999), the same source used for the original items. The first cohort of 21 children were administered the 20-item version of these measures (described in the Method section) during their follow-up year, whereas the 48 children in the second cohort (25 treatment; 23 control) received the 20-item version during both years of the study. To make the scores of the 21 children in the first cohort who received the 15-item version of these measures during the treatment year comparable to their scores on the 20-item version administered during their follow-up year (and the 20-item version administered to all 48 children in the second cohort), we prorated the scores of the 21 children during the treatment year so that they range from 0–20.

replaced any remedial reading instruction that might otherwise have been provided by the school for that year (e.g., Chapter 1 reading, resource room instruction in reading for children identified by their school as having a disability). Taking into account field trips, absences, and other unexpected events, tutored students participated, on average, in 126 tutoring sessions (range = 103–138 sessions) or 105 hr of tutoring (range = 86–115 hr). Each child continued to get regular classroom reading instruction. Classroom teachers of the treatment children provided descriptions of their reading programs in the spring of the treatment year. A majority of these teachers described their program as literature based, with some emphasis on phonics.

Tutor Qualifications and Training

Tutoring was provided by 12 teachers certified in either reading or special education. All tutors, except 2, had completed a master's degree in one of these areas. Both of the tutors who had not yet completed their master's degrees were dually certified in elementary education and special education and were enrolled in graduate programs (in special education and reading, respectively). Tutors were assigned to treatment children on the basis of logistical considerations. That is, when possible, tutors were assigned to schools that were more convenient for them in terms of location, hours of operation, number of students to be tutored (e.g., 3 versus 4). Prior to working with the treatment children, tutors completed a 45-hr, 15-session training program in August and September, directed by a colleague with extensive early intervention and teacher training experience. The training program included a review of the project; current research in early reading acquisition, with an emphasis on phonological processing and alphabetic coding; the structure of the English language; and specific practice using the systematic and explicit teaching strategies to be followed in this study. Tutors also met with Benita A. Blachman and Sheila M. Clonan for eight additional 2-hr training sessions—approximately one each month beginning in October. Initially, the focus of these sessions was on a review of the tutoring protocol (the five parts of the lesson required each day), feedback from observations of individual tutors that might have relevance to all tutors (e.g., concerns regarding pacing), procedures to build fluency (e.g., timed reading of individual words, rereading books), review of the developmental sequence of the program and strategies to help children transition from one level to the next based on learning each of six syllable types, and integration of word-level work and text-based reading. Later sessions included strategies for introducing multisyllable words, a greater focus on books appropriate for children whose word-level skills were improving (e.g., chapter books, nonfiction texts), and strategies for building comprehension (e.g., summarizing, inferencing).

Treatment Group Instruction

The intervention was based on a framework for organizing instruction that had been used previously in prevention studies (Blachman, 1987; Blachman, Tangel, Ball, Black, & McGraw, 1999) and was adapted in this study for a slightly older population (second and third graders) of remedial students. The tutoring program included explicit and systematic instruction to help children develop an understanding of the phonologic and orthographic connections in words, as well as many opportunities to read texts that were phonetically controlled and texts that were not phonetically controlled (both narrative and expository texts) to develop fluency, build comprehension strategies, and foster reading for pleasure. Each lesson was built around a five-step plan and extended activities. Although tutors were expected to include each of the five steps in each lesson (and the inclusion of each step was the basis for monitoring treatment fidelity), the program was not scripted. Each lesson was individualized based on the child's progress and on feedback provided to the tutor from observations by Benita A. Blachman and Sheila M. Clonan and the tutor coordinator (a doctoral student in reading with training in the intervention, who rotated among the 11 schools to provide on-site supervision to the tutors). The five steps in the lesson are described below.

1. Each lesson began with a brief and quick-paced review of sound–symbol associations learned in previous lessons and the introduction of new sound–symbol correspondences. For this part of the lesson, tutors used a sound pack (set of index cards) containing each of the graphemes (i.e., letters and letter clusters) being reviewed. To draw attention to the vowels, we printed the vowels (e.g., *a*) and, later, vowel combinations (e.g., *ai*) in red.

2. Next, each child practiced phoneme analysis and blending by manipulating letter cards on a sound board or using Scrabble tiles to make new words reflecting a particular syllable pattern (e.g., closed syllables, such as *hid*; final "e" syllables, such as *hide*). This followed a very systematic sequence. For example, in an early lesson the child might be asked to manipulate letters to change *fan* to *fat* to *sat* to *sag*. When new vowels were mastered, the child would learn to change *fan* to *fin* to *shin*. As new syllable types were introduced, the child might be asked to change *shin* to *shine* and, later, to combine syllable types previously learned, making the word *sunshine*. As students worked through this step in the lesson, teachers also asked questions about the words the children were producing, such as "How many sounds? How many letters? What happens to *shin* if we add an *e* at the end? What happens if we take it away?"

3. A fluency building activity was introduced next. This activity provided an opportunity for each child to develop more automatic recognition of syllable patterns that they had practiced previously on the sound board. Once the child could construct and accurately read a pool of phonetically regular words, these words (and previously unintroduced words with similar patterns) were put on flash cards and the child practiced reading them quickly. High frequency irregular words, such as *said*, were selected from the Instant Word lists compiled by Fry, Kress, and Fountoukidis (1993) and introduced at this time. These words were written in a different color than the phonetically regular words. For several minutes each day, children practiced reading both phonetically regular words and high frequency words. Tutors generally reviewed the lists more than once, timing the child so speed could be compared within a single session or across several days.

4. The fourth step in the lesson consisted of oral reading practice, during which children read phonetically controlled text (e.g., *Primary Phonics* [Makar, 1995]; *Steck-Vaughn Phonics Readers* series from Steck-Vaughn); trade books that were not phonetically controlled (e.g., *Amelia Bedelia* series by Peggy Parish, *Arthur* series by Marc Brown); and expository texts reflecting, most often, science themes (e.g., *Curious Creatures* series from Curriculum Associates). Trade books were selected on the basis of appropriateness of the reading level and on the child's interests; therefore, the selections varied across children.

5. In the final required element of each lesson, tutors dictated words used in earlier steps of the lesson (e.g., words practiced on the sound board or encountered in phonetically controlled text) or new words with the same phonetic pattern. Generally, six to eight phonetically regular words and two sentences were dictated. The child was directed to print vowel headings at the top of each dictation page (e.g., *a* and *i*, or later in the year, *ai, oa, ea*). These headings included some of the vowel graphemes that represented the target sounds for that day's lesson. The dictation activity gave tutors an opportunity to evaluate student progress, and the dictation notebooks became a record of student growth over the year, as students progressed from writing and reading simple closed syllable words (e.g., *ham*) to more complex syllable types (e.g., *hike, rain, starch*) and multisyllable words (e.g., *reptile, bugle, tarnish*) made up of the syllable types they had learned. Extended activities were also included in the last 5–10 min of the lesson, if time permitted. These activities included additional text-based reading, journal writing, or games to reinforce previously learned skills.

The program was divided into six levels, but because not all children progressed at the same rate, not all children completed the program. Those who completed the program were introduced to words representing all six syllable types, including closed (e.g., *fat*, *flat*), final "e" (e.g., *cake*, *shine*), open (e.g., *me*, *cry*), vowel team (e.g., *pain*, *teach*, *coin*), vowel + r (e.g., *burn, start*), and consonant-"le" syllables (e.g., bot*tle*, ta*ble*), as well as multisyllable words combining these syllable patterns. Although the focus of many of the early activities was on developing accurate and fluent word recognition skills, tutors were encouraged to make sure that each child knew the meaning of all words that they were asked to read or spell, and comprehension of stories was developed through a variety of strategies (e.g., summarizing, making inferences). By expanding an earlier version of this program from 30 min to 50 min, we were able to increase the sophistication of the lessons, primarily by including more work on reading multisyllable words, timed reading of lists of words, and additional time on spelling and text-based reading.

Treatment Fidelity

To monitor treatment fidelity, each child was observed an average of nine times during the treatment year by Benita A. Blachman, Sheila M. Clonan, or our tutor coordinator. On the basis of an analysis of these observation protocols by two independent raters, 96% of the observed lessons included all five of the required steps, with 100% interrater agreement. In addition, tutors were required to turn in a two-page lesson plan for each tutoring session for each child and to turn in one audiotaped lesson each week for each child they tutored. The observations and reviews of audiotapes were used to ensure treatment fidelity and as a vehicle to discuss the progress of each child with the tutor. As a final measure of treatment fidelity, two tapes for each child (one from a 3-week period in the fall and one from a 3-week period in late spring) were reviewed by two independent reviewers after the treatment year ended. On the basis of these reviews, 90% of the tapes (67 of 74 tapes reviewed) included all five steps of the lesson. There was 100% interrater agreement regarding adherence to the treatment protocol. When we looked more carefully at the 7 tapes with missing steps, we found that on 2 of the 7 tapes the tape recorder had been turned off prior to either the sound board or dictation portion of the lesson. Because the lesson plan turned in by the tutor for the taped lesson included notes regarding the child's performance on the missing step, we concluded that the step was actually included but not taped. Consequently, 90% compliance is likely to be an underestimate of treatment fidelity.

Control Group Instruction

Control children continued to receive whatever remedial reading instruction was provided by the school (e.g., Chapter 1 reading, reading instruction in the resource room), in addition to their classroom reading instruction. Nine of the 32 control children received no special reading help outside the regular classroom during the treatment year. This was because prior to the start of the school year, 2 of the 11 participating schools eliminated all of the reading teachers in their district in an effort to reduce class size. That is, reading teachers could keep their jobs if they were willing to become classroom teachers. Eight of the 9 control children who received no special reading help outside the classroom attended these schools. To examine the differences between the 9 children who did not receive treatment and the 23 control children who did, we compared the reading and spelling pretest scores for these two groups on both standardized (WRMT–R Basic Skills Cluster, GORT–3 Quotient, and WRAT3 Spelling) and nonstandardized (Word Reading, Word Reading Efficiency, and Spelling Dictation) measures. We found no significant pretest differences between these groups on any of the reading and spelling measures. In addition, we also examined whether these two groups of control children exhibited posttest differences in reading and spelling after controlling for pretest reading and spelling scores. The results of these analyses revealed no differences in change from pretest to posttest for these two groups on any of these measures. On these analyses, the control children who did not receive remedial help during the treatment year were not different from the control children who received additional help in reading during the treatment year. It is also worth noting that even though the two groups of control children were not statistically different from each other at pretest or at posttest, the children who received no help gained more standard score points on several measures than those who received help, although the differences were not statistically significant.

The 23 control children who received remedial reading instruction participated in small groups (averaging 4 children per group and ranging from 2 to 8) that met outside the regular classroom from three to five times per week for an average of 41 min per session (range = 25–90 min). Control children participated in an average of 104 sessions (range = 58–144 sessions) or an average of 77 hr (range = 29–212 hr) of remedial reading instruction between the pretests administered at the end of September and the posttests administered at the beginning of June. The reading and resource teachers who provided remedial reading instruction to the control children were all certified in either reading or special education, and all had a master's degree in one of these areas. These teachers were almost equally divided in how they characterized their programs. Slightly more than half characterized their programs as phonics based, with some literature, and somewhat less than half characterized their programs as literature based, with some phonics. Descriptions of classroom reading programs were also provided by the regular classroom teachers of the control children. The majority of these teachers described their programs as primarily literature based, with some phonics.

Follow-Up Year

For 1 year following the treatment, the progress of all children was monitored by means of the skills battery administered four times during the school year, in September, December, March, and June. In June of the follow-up year, the original pretest battery was also administered again (including the Woodcock Word Identification and Word Attack measures originally used for screening purposes). As indicated previously, all testing during both the treatment year and the follow-up year was administered by trained examiners blind to the condition of the children. It is worthy of note that there was no attrition during the study. We continued to follow all 69 children during the follow-up year, assessing children who moved from one school to another within their original school districts, as well as following students who moved to a different school district.

During the follow-up year, all children received regular classroom reading instruction and, in some cases, were assigned a reading or resource teacher by the school to provide remedial reading instruction outside the classroom. We again collected data from the reading and resource teachers regarding the remedial reading programs for each child in the study and also collected data from classroom teachers regarding their general characterization of their classroom reading program. Fifty-one percent (19 of 37 children) of the treatment children and 63% (20 of 32 children) of the control children received special reading instruction outside the classroom during the follow-up year. Remedial reading instruction for both groups consisted of three to five small group sessions each week for an average of 45 min per session (with a range of 30–75 min). Children participated in groups ranging in size from 2 to 8, with an average of 4 per group, with two exceptions–1 treatment child was in a group of 12 and 1 control child received individual tutoring. Children who had been in the treatment group participated, on average, in 120 sessions (range = 41–144), or an average of 93 hr (range = 31–143) of remedial reading instruction during the follow-up year, and the control children participated, on average, in 127 sessions (range = 72–144), or an average of 97 hr (range = 36–144) of remedial reading instruction during the follow-up year. There was no difference in the number of hours of instruction provided to the treatment and control children who received remedial reading instruction during the follow-up year, $t(37) = 0.40$, $p = .69$, and no difference in the way the remedial reading teachers of each group characterized their programs (Fisher's exact test, $n = 39$, $p = .42$), with the majority reporting their instruction to be phonics based, with some literature. The majority of the regular class teachers of both the treatment and the control children characterized their classroom reading programs as primarily literature based, with some emphasis on phonics.

Results

Analyses

In order to investigate the effects of the experimental tutoring condition, we used two approaches. First, we looked at mean level comparisons between the treatment group and the control group as a whole during the treatment year and follow-up year. For the standardized (norm-referenced) measures that were collected at three time points (pretest, posttest, follow-up), the data were analyzed with an analysis of covariance (ANCOVA) approach that covaried the pretest measurement from the posttest and follow-up measurements. In analyzing randomized experimental designs, this technique has been found to be more powerful than a repeated measures analysis of variance (ANOVA) approach (Maxwell, 1998).

For those nonstandardized (nonnormed) measures that were administered four times during the treatment year and four times during the follow-up year, we used an individual growth curve approach in which we estimated growth and overall level of performance for each individual during the treatment year and follow-up year (and allowed for different rates of growth and overall levels of performance across the 2 years). This method also provides for an estimation of an average growth rate and level of performance for each group.

In a second set of analyses designed to further explore the effects of treatment, we divided both the treatment group and the control group into two ability groups (a higher and a lower skilled treatment group and a higher and lower skilled control group) on the basis of their initial performance on the Woodcock Basic Skills Cluster. Analyses on these groups were conducted with a series of 2×2 ANCOVAs comparing the posttest means of the treatment and control groups and the higher initial skill and lower initial skill groups, while covarying the pretest measurement. These analyses were conducted to investigate possible differential treatment effects based on initial level of ability by comparing treatment children to control children with comparable levels of initial performance. For example, it was possible that the treatment was more effective for the children who came into the study with higher levels of reading ability than for those children with lower levels.

Treatment Group Effects

Treatment effects were estimated by means of the ANCOVA approach previously described. Even though children were randomly assigned to the treatment and control conditions, we inspected the pretest means to investigate possible pretreatment differences between the treatment and control groups. These results appear in Table 1.[3] The pretest means were analyzed with a series of *t* tests, and the posttest and follow-up means were covaried by that measure's respective pretest score, with the adjusted means also reported. The *F* tests associated with these mean comparisons test for the differences between the treatment and control means, adjusted for the pretest measurement. We calculated effect sizes by first covarying the pretest score from the dependent variable and using these residualized (adjusted) means and standard deviations to estimate the size of the effect. These effects are reported in standard deviation units (Cohen's *d*; Cohen, 1988).

As can be seen from Table 1, there were no differences between the treatment group and the control group at pretest for any of the standardized measures (as well as no difference in age at pretest between the two groups). At posttest, however, there were significant differences between the treatment and control groups on all of the reading and spelling measures, with effect sizes ranging from 1.69 for the WRMT–R Basic Skills Cluster to 0.55 for GORT–3 Comprehension. In contrast, smaller effect sizes on two measures of mathematical ability (WJ–R Calculations and Applied Problems subtests) provide some evidence for the discriminant validity of the treatment effects.

Results for the follow-up year revealed a similar pattern of findings, although some of the differences that were significant at posttest were no longer significant at follow-up. Significant differences between the groups remained at follow-up on all of the measures of reading and spelling except for two subtests of the GORT–3 (Accuracy and Comprehension).

Analyses of the nonstandardized measures were conducted using growth curve analysis (Bryk & Raudenbush, 1992; Francis, Schatschneider, & Carlson, 2000). Specifically, growth on measures of Word Reading, Word Reading Efficiency, and Spelling was modeled, along with three measures of phonological processing (phonological awareness, rapid naming of letters, and nonword repetition), using SAS PROC MIXED V8.1 (SAS Institute, 2001). These models were constructed so that estimates of linear growth and overall performance on these measures were estimated for both the treatment year and follow-up year. Post hoc contrasts were used to examine differential growth in both years of the study for the two comparison groups, as well as comparisons of overall mean level performance at the end of the treatment year and at the end of the follow-up year for the two groups. The raw means and standard deviations for these six variables are presented in Table 2 for both the treatment year and the follow-up year, and the predicted means and growth rates for these measures for the treatment year and follow-up year are presented in Table 3. An analysis of possible pretest differences on these variables found that the

[3] Recognizing the arbitrariness of statistical significance (Wainer & Robinson, 2003), we provide the computed *p* values so that readers can apply their own criterion for judgment of statistical significance. Interpretation of the results is not altered in any substantial way if readers elect to apply an adjustment for multiple tests. That is, the intervention had an effect on reading and spelling, with the smallest effect on reading comprehension. There is considerable debate about the best approach when adjusting for multiple tests (Holland & Copenhaver, 1988; Seaman, Levin, & Serlin, 1991). If we take a conservative approach and adjust by the total number of reading and spelling tests, the *p* value for significance would be .006 for the treatment year. This approach is considered by some (e.g., Jaccard & Wan, 1996) to be too conservative, and alternative procedures have been suggested to reduce the chance of a Type II error (e.g., Uitenbroek, 1997), including a modified Bonferroni method based on Holm (1979) that is considered "more powerful than traditional Bonferroni methods but adequately maintains experimentwise error rates at the desired alpha level (usually 0.05)" (Jaccard & Wan, p. 30). In this modified approach, the observed *p* values are rank ordered, the smallest appearing first. The alpha level required for significance for the smallest *p* value would be $.05/c$ (or .006, where $c = 8$, the number of posttreatment reading and spelling tests). If the null hypothesis is rejected, the next smallest *p* value is judged against $.05/(c - 1)$, or .007, the next smallest against $.05/(c - 2)$, or .008, and so on. Applying this correction procedure did not alter the results.

Table 2
Means and Standard Deviations for Treatment[a] and Control[a] Groups for the Nonstandardized Measures Given Four Times During the Treatment Year and Four Times During the Follow-Up Year

Task and group	September *M*	September *SD*	December *M*	December *SD*	March *M*	March *SD*	June *M*	June *SD*
Treatment year								
Word Reading[b]								
Treatment	9.11	8.27	14.51	9.87	20.89	10.89	26.14	10.11
Control	7.66	8.57	10.97	9.40	16.47	10.59	18.44	11.56
Word Reading Efficiency[c]								
Treatment	0.43	0.22	0.59	0.23	0.75	0.26	0.85	0.25
Control	0.38	0.21	0.47	0.25	0.58	0.26	0.65	0.29
Spelling[d]								
Treatment	5.05	5.00	8.68	6.14	13.70	8.05	17.14	9.09
Control	3.41	4.48	5.78	5.68	8.75	7.47	10.59	7.53
Phonological Awareness[e]								
Treatment	30.14	12.55	41.46	10.50	47.61	11.24	49.45	10.07
Control	31.27	13.22	39.28	13.93	43.43	12.74	43.73	14.75
Rapid Naming of Letters[f]								
Treatment	1.21	0.30	1.48	0.26	1.64	0.27	1.61	0.36
Control	1.17	0.35	1.43	0.36	1.50	0.35	1.53	0.39
Nonword Repetition[g]								
Treatment	14.00	3.33	14.43	4.65	15.46	4.54	16.54	4.43
Control	12.19	3.95	14.00	4.41	14.56	4.11	15.91	3.98
Follow-up year								
Word Reading[b]								
Treatment	27.73	8.54	30.08	8.18	32.08	7.08	33.05	6.35
Control	20.38	11.09	22.09	11.28	25.16	10.57	26.78	10.23
Word Reading Efficiency[c]								
Treatment	0.87	0.23	0.95	0.23	1.02	0.22	1.06	0.24
Control	0.66	0.29	0.74	0.29	0.81	0.28	0.86	0.31
Spelling[d]								
Treatment	17.00	8.79	20.14	9.22	22.49	10.49	23.27	10.51
Control	10.88	8.09	13.00	7.87	15.41	10.12	17.81	10.33
Phonological Awareness[e]								
Treatment	48.99	14.14	52.98	12.43	54.58	13.95	56.24	14.01
Control	45.02	15.59	50.20	12.73	51.54	12.80	51.52	13.48
Rapid Naming of Letters[f]								
Treatment	1.60	0.29	1.78	0.32	1.81	0.30	1.88	0.39
Control	1.56	0.38	1.68	0.45	1.76	0.51	1.75	0.48
Nonword Repetition[g]								
Treatment	15.81	3.89	16.59	4.22	16.43	4.25	17.84	3.95
Control	15.59	4.32	16.00	4.22	17.75	4.12	17.47	4.27

[a] Treatment group $n = 37$; control group $n = 32$. [b] Word Reading raw score range 0–41. [c] Word Reading Efficiency in words per second. [d] Spelling raw score range 0–50. [e] Phonological Awareness raw score range 0–75. [f] Rapid Naming of Letters in letters per second. [g] Nonword Repetition raw score range 0–25.

groups did differ significantly on nonword repetition, with the difference favoring the treatment group, $t(67) = 2.07$, $p = .0425$. The groups did not differ significantly on the other five measures at pretest.

With regard to the reading and spelling measures, at the end of the treatment year, significant differences between the groups were found on Word Reading, $t(416) = 3.70$, $p = .0002$; Word Reading Efficiency, $t(416) = 4.87$, $p < .0001$; and Spelling, $t(416) = 3.79$, $p = .0002$. Differential growth rates during the treatment year, favoring the treatment group, were also found on all three measures: Word Reading, $t(416) = 2.69$, $p = .0074$; Word Reading Efficiency, $t(416) = 2.67$, $p = .0079$; and Spelling, $t(416) = 2.55$, $p = .0110$. These results are consistent with the findings from the standardized tests, indicating increased gains in these skills for the treatment group as compared with the control group. For the end of the follow-up year means, differences between the groups remained significant on all three measures: Word Reading, $t(416) = 2.72$, $p = .0069$; Word Reading Efficiency, $t(416) = 4.24$, $p < .0001$; and Spelling, $t(416) = 2.75$, $p = .0063$. The rates of growth, however, for the two groups during the follow-up year were not significantly different on any of the three measures. Thus, although during the follow-up year the treatment group maintained their gains, their rate of growth during the follow-up year did not differentially increase or decrease relative to the control children. Graphs depicting this growth appear in Figures 1–3.

Table 3
Predicted Means and Growth Rates for Treatment[a] and Control[a] Groups

Task and group	Treatment year: Predicted mean at end of year	Treatment year: Predicted rate of growth[b]	Follow-up year: Predicted mean at end of year	Follow-up year: Predicted rate of growth[b]
Word Reading				
Treatment	25.55**	16.71**	32.65**	5.12
Control	20.09	11.38	28.65	7.62
Word Reading Efficiency				
Treatment	0.86**	0.41**	1.06**	0.18
Control	0.68	0.28	0.90	0.22
Spelling				
Treatment	16.57**	11.79*	23.09**	5.82
Control	11.44	7.22	19.38	8.30
Phonological Awareness				
Treatment	51.70*	17.64*	56.57	6.27
Control	47.20	11.79	54.50	7.19
Rapid Naming of Letters				
Treatment	1.68*	0.40	1.89*	0.24
Control	1.55	0.34	1.75	0.16
Nonword Repetition				
Treatment	16.17	2.51	17.25	1.72
Control	15.67	3.28	17.77	2.31

[a] Treatment group $n = 37$; control group $n = 32$. [b] Growth rate unit for Word Reading, Spelling, Phonological Awareness, and Nonword Repetition is number correct per school year (September to June). Growth rate for Word Reading Efficiency and Rapid Naming of Letters is number correct per second per school year (September to June).
* $p < .05$, significant within-year difference between treatment and control. ** $p < .01$, significant within-year difference between treatment and control.

For the three phonological processing measures, there were significant differences on two of the three measures at the end of the treatment year; specifically, there were differences on phonological awareness, $t(416) = 2.22$, $p = .0268$, and rapid naming of letters, $t(416) = 2.15$, $p = .0322$. Significant differences between groups in growth rates during the treatment year were also found for phonological awareness, $t(416) = 2.06$, $p = .0404$. All of these differences favored the treatment group. Significant group differences at the end of the follow-up year remained only for rapid naming of letters, $t(416) = 2.27$, $p = .0235$. The groups were no longer significantly different on phonological awareness. As was the case with the reading outcome measures, the rates of growth during the follow-up year were not significantly different on any of the three phonological processing measures, indicating that the growth on these measures was the same for both the treatment and control groups during the follow-up year.

Treatment × Initial Skill Analyses

As indicated previously, to further explore the nature of the treatment effects, we divided each group into two groups on the basis of their pretest scores on the Woodcock Basic Skills Cluster and conducted a series of 2 (treatment groups) × 2 (initial skill groups) ANCOVAs on a subset of the standardized battery. Initial skill groups were formed by placing children who started the year above the 15th percentile on the WRMT–R Basic Skills Cluster into a higher initial skill group, and children who started the year at or below the 15th percentile on the WRMT–R Basic Skills Cluster into a lower initial skill group. After the groups were formed, the higher initial skill treatment group had a mean Basic Skills Cluster percentile at pretest of 19.21 ($SD = 2.62$, $n = 19$), and the higher initial skill control group had a mean percentile score at pretest of 20.25 ($SD = 2.62$, $n = 16$). For the lower initial skill treatment group, the mean percentile score at pretest was 7.28 ($SD = 4.84$, $n = 18$), and the lower initial skill control group had a mean percentile score at pretest of 7.69 ($SD = 3.82$, $n = 16$). Planned contrasts were performed, with each ANOVA comparing the treatment and control children within each initial skill level group. The results of the planned comparisons are presented in Table 4 for the children in the higher initial skill group, and in Table 5 for the children in the lower initial skill group. Following the analysis of the overall treatment effect, the posttest and follow-up means were covaried by that measure's respective pretest score, with the adjusted means also reported. The F tests associated with these mean comparisons, test for the differences between the treatment and control means, adjusted for the pretest. Effect sizes are also reported. Although the data are presented in two separate tables, it is worth noting that there was only one analysis per variable for the treatment year and follow-up year, respectively.

Although there were no differences between the higher initial skill treatment group and the higher initial skill control group at pretest, there were significant differences between groups on the WRMT–R Basic Skills Cluster and WRAT3 Spelling at the end of the treatment year and again at the end of the follow-up year, with these differences favoring the higher initial skill treatment group. There was no significant difference, however, between these two groups on the GORT–3 Quotient at either posttest or follow-up.

The comparison of the lower initial skill treatment group to the lower initial skill control group revealed a somewhat similar

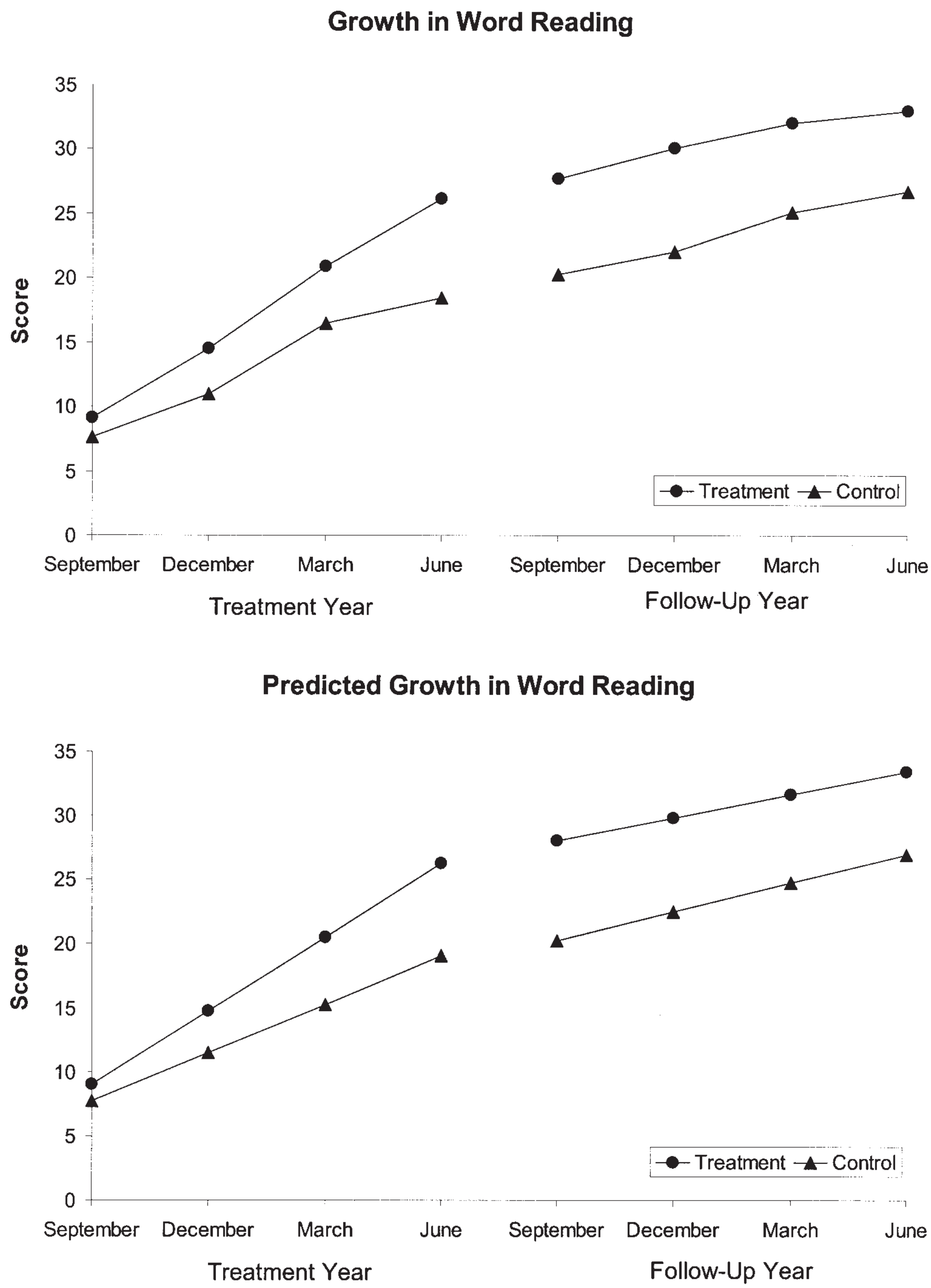

Figure 1. Growth and predicted growth in word reading.

pattern of results. Again, there were no pretest differences between groups. However, at the end of the treatment year and at the end of the follow-up year, the lower initial skill treatment group differed significantly from the lower initial skill control group on the WRMT–R Basic Skills Cluster and WRAT3 Spelling. In contrast to the higher initial skill group comparisons, these two lower initial skill groups also differed significantly on the GORT–3 Quotient at both posttest and follow-up, with the difference favoring the group of lower initial skill children who received treatment.

Discussion

Our primary goal in this study was to evaluate an intervention for second- and third-grade students with poor word-level skills. The major findings are that the treatment children, who participated in an intensive, systematic, and explicit program that emphasized the phonologic and orthographic connections in words and text-based reading, showed significantly greater gains than the control children on measures of both real word and nonword

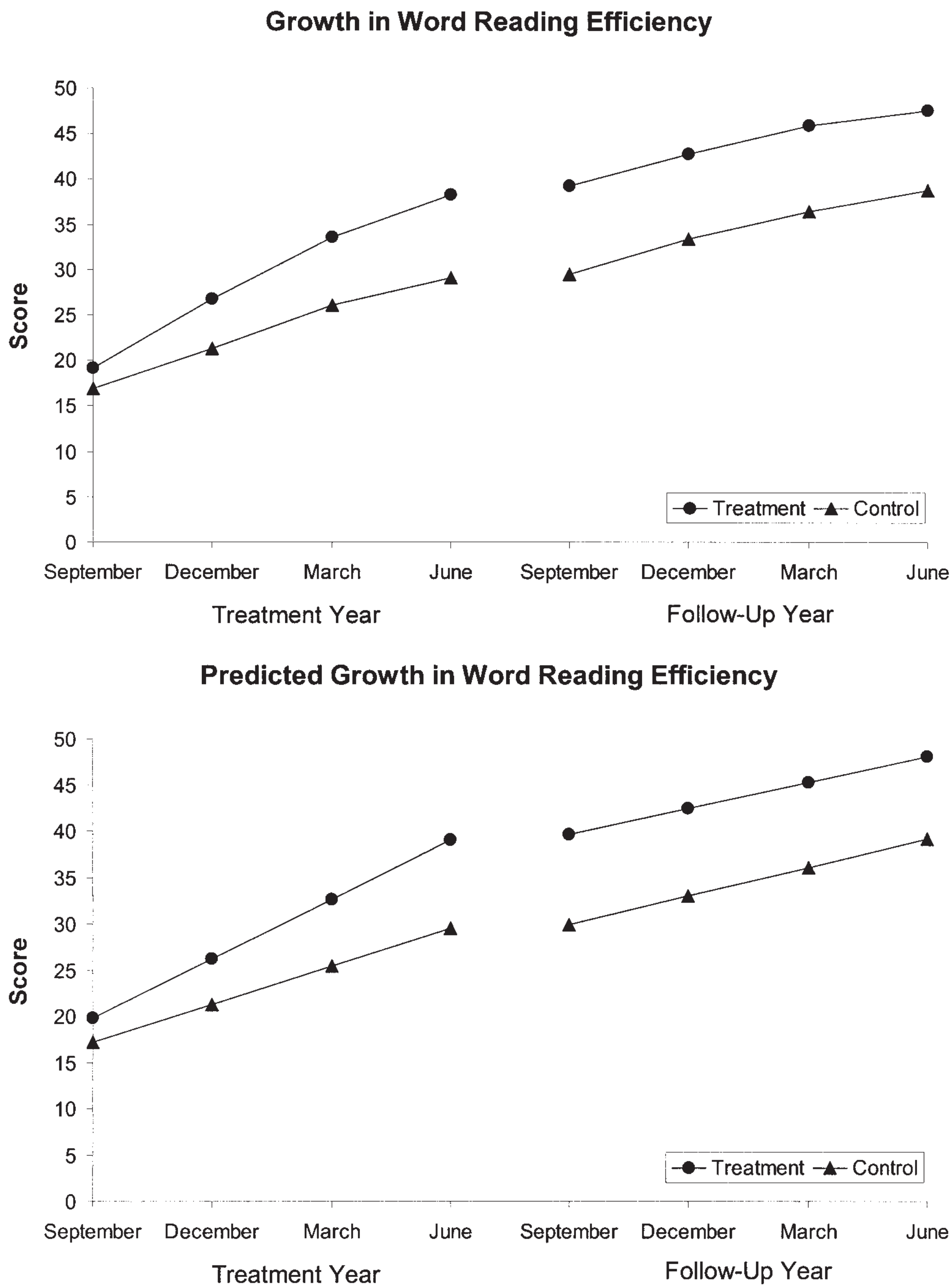

Figure 2. Growth and predicted growth in word reading efficiency.

reading, reading rate, passage reading, and spelling and, for the most part, maintained these gains at a 1-year follow-up. As further evidence that our intervention specifically targeted reading, as planned, there were no group differences in math at either pretest or posttest.

Although others have not always found transfer effects from programs that emphasize phonology to real word reading (as opposed to nonword reading; Byrne & Fielding-Barnsley, 1995), our treatment children demonstrated significant gains in real word reading on both standardized and nonstandardized measures at posttest and 1 year later. It is possible that our treatment provided more balance than some other programs, in that children were explicitly introduced to both phonetically regular and exception words (although by far the preponderance of isolated word work was spent on learning to read and spell phonetically regular words representing the six syllable patterns in English). Drawing on Ehri's (1997) theory of sight word learning, transfer from phonetically based instruction to improved reading of irregular as well as regular words should not be surprising. Ehri postulates that the process of forming connections between

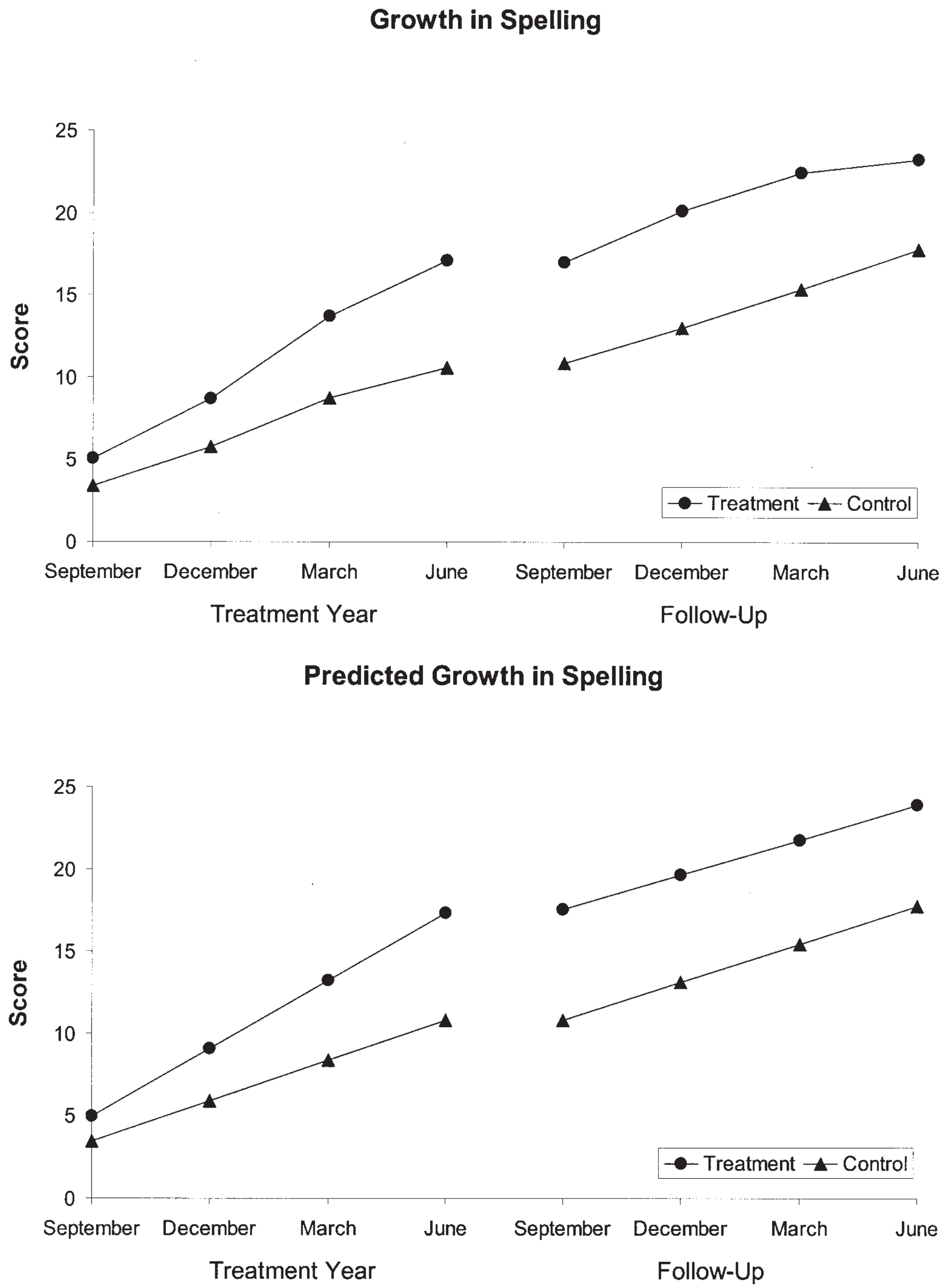

Figure 3. Growth and predicted growth in spelling.

graphemes and phonemes facilitates storing words in memory and that this process

> allows readers to remember how to read not only words containing conventional grapheme-phoneme correspondences, such as *stop*, but also words that are spelled irregularly. . . . It turns out that most of the letters in irregular words conform to grapheme-phoneme conventions, for example, all but S in *island.* (Ehri, 1997, p. 171)

Another area in which we found treatment effects at both posttest and follow-up was spelling. This is an important finding in light of the recent meta-analysis conducted by the NRP (2000), which revealed that systematic phonics instruction did not significantly improve the spelling of children who were poor readers above first grade, as evidenced by an effect size across studies of 0.09. In contrast, our tutoring intervention resulted in large effect sizes for spelling of 1.13 at the end of the treatment year and 0.81 at the end of the follow-up year. The significant group differences in spelling in this study are also in contrast to results from a classroom prevention study (Blachman et al., 1999), in which we used an earlier version of the intervention reported here. In the

Table 4
Pretest, Posttest, and Follow-Up Means and Adjusted Means for Higher Initial Skill Treatment and Control Groups

Measure	Treatment[a]			Control[a]					
	M	*SD*	*M*[b]	*M*	*SD*	*M*[b]	*F*	*p*	*d*[c]
Pretest only									
Age at entry (years)	7.93	0.57		7.66	0.52		2.07	0.1583	
WISC–III Verbal IQ	96.68	9.62		93.81	9.46		0.79	0.3818	
WISC–III Performance IQ	99.74	11.06		102.25	14.49		0.34	0.5647	
WISC–III Full Scale IQ	97.79	9.25		97.44	12.18		0.01	0.9232	
Pretest									
WRMT Basic Skills Cluster	86.95	1.31		87.31	1.49		0.59	0.4465	
WRAT Spelling	86.21	6.01		84.81	4.89		0.55	0.4616	
GORT Quotient	77.74	7.70		75.81	6.79		0.61	0.4430	
Posttest									
WRMT Basic Skills Cluster	95.42	9.93	95.74	85.56	7.27	85.18	13.10	0.0010	1.25
WRAT Spelling	96.84	7.20	96.78	87.75	9.56	87.82	9.55	0.0041	1.04
GORT Quotient	90.21	10.19	89.99	84.06	9.28	84.33	2.86	0.1000	0.58
Follow-up									
WRMT Basic Skills Cluster	94.37	9.54	94.66	87.19	7.99	86.84	6.92	0.0129	0.90
WRAT Spelling	97.05	7.47	97.05	89.31	5.82	89.31	10.82	0.0024	1.13
GORT Quotient	89.89	9.71	89.39	86.13	8.33	86.72	0.90	0.3497	0.33

Note. WISC–III = Wechsler Intelligence Scale for Children—Third Edition; WRMT = Woodcock Reading Mastery Tests—Revised; WRAT = Wide Range Achievement Test 3; GORT = Gray Oral Reading Tests—Third Edition.
[a] Higher initial skill treatment group $n = 19$; higher initial skill control group $n = 16$. [b] Means adjusted by pretest. [c] d = Cohen's measure of effect size.

earlier study, we found differences in spelling on a standardized measure at the end of first grade, but no differences on the same measure 1 year later. One of the modifications we made when we adapted our classroom intervention for use with remedial students during tutoring was to put more emphasis on spelling. In the classroom study, teachers did not always include spelling (the last step of the lesson) or often alternated spelling with sound board activities in order to complete the lesson within the 30 min allocated for the intervention. In the study reported here, dictation was included daily, and children who completed the program, and who were thus exposed to all six syllable patterns, had the opportunity to practice writing more complex words during dictation (e.g., *forgave*, *wiggle*). As is frequently the case in multistep programs, we were not able to identify with certainty which component of the program was most responsible for a particular outcome (e.g., spelling). We can, however, speculate that the increased emphasis placed on spelling in this study contributed to the gains in spelling among the treatment children.

Another area that, like spelling, has been found to be difficult to influence in even the most effective interventions is reading rate (Torgesen et al., 2001). At the end of treatment and 1 year later, our treatment children showed significant gains over the control children on three measures of rate—a nonstandardized measure of rapid naming of letters, a nonstandardized measure of word reading efficiency, and the standardized reading rate measure on the GORT–3, with effect sizes on the latter measure of 0.96 at posttest and 0.81 at follow-up. Again, given the multiple components in our intervention, it is not clear which instructional component had a direct effect on reading rate. However, as with spelling, one of the modifications we made when we adapted our classroom model for the remedial students in this study was to put more emphasis on building automatic word recognition by including timed reading of isolated word lists. We did not, however, include timed reading of text. Although some research suggests that rate improves regardless of whether children practice reading words in lists or in text (Levy, 2001), an unanswered question is whether gains in this area would have been even stronger if timed reading of text had also been included.

Although our students differed on our one measure of comprehension at the end of the treatment year, these differences were not maintained at the end of the follow-up year when the treatment children were no longer exposed to the experimental intervention. Recent evidence suggests that explicit instruction can improve not only the word reading and spelling skills of poor readers, but comprehension as well (RAND Reading Study Group, 2002; Williams et al., 2002). This was the area in which our program was least explicit and systematic. That is, although tutors were encouraged to have students practice strategies found to contribute to improved comprehension (Dole, Duffy, Roehler, & Pearson, 1991), such as summarizing and drawing inferences, much less guidance was provided for the tutors in this area, and the degree to which these strategies were included in each lesson was not

Table 5
Pretest, Posttest, and Follow-Up Means and Adjusted Means for Lower Initial Skill Treatment and Control Groups

Measure	Treatment[a]			Control[a]					
	M	*SD*	*M*[b]	*M*	*SD*	*M*[b]	*F*	*p*	*d*[c]
Pretest only									
Age at entry (years)	7.98	0.41		7.99	0.59		0.00	0.9555	
WISC–III Verbal IQ	92.83	10.42		91.94	9.08		0.07	0.7922	
WISC–III Performance IQ	98.56	15.81		94.69	13.18		0.59	0.4475	
WISC–III Full Scale IQ	94.89	11.91		92.81	10.46		0.29	0.5949	
Pretest									
WRMT Basic Skills Cluster	76.56	6.55		77.44	5.07		0.19	0.6665	
WRAT Spelling	77.72	3.54		78.06	7.19		0.03	0.8653	
GORT Quotient	68.17	5.06		72.25	6.97		3.88	0.0574	
Posttest									
WRMT Basic Skills Cluster	80.83	10.78	81.33	72.19	6.48	71.63	23.62	0.0001	1.71
WRAT Spelling	87.33	6.95	87.42	77.88	6.76	77.78	19.89	0.0001	1.57
GORT Quotient	79.33	9.76	80.15	71.50	6.84	70.58	10.18	0.0032	1.04
Follow-up									
WRMT Basic Skills Cluster	80.39	9.27	80.64	71.94	8.68	71.65	9.67	0.0040	1.07
WRAT Spelling	83.78	7.38	83.90	77.81	8.89	77.67	6.81	0.0137	0.91
GORT Quotient	79.50	11.33	81.22	73.75	12.56	71.81	5.71	0.0229	0.78

Note. WISC–III = Wechsler Intelligence Scale for Children—Third Edition; WRMT = Woodcock Reading Mastery Tests—Revised; WRAT = Wide Range Achievement Test 3; GORT = Gray Oral Reading Tests—Third Edition.
[a] Lower initial skill treatment group $n = 18$; lower initial skill control group $n = 16$. [b] Means adjusted by pretest. [c] d = Cohen's measure of effect size.

monitored. An important question for future research is whether including more explicit and systematic comprehension instruction during the treatment year would have resulted in gains in comprehension that were maintained during the follow-up year.

In addition to our interest in differences in reading and spelling, we were also interested in rate of growth during both years of the study. During the treatment year, the intervention resulted in an accelerated growth trajectory for the treatment children compared with the control children, providing additional evidence that when children receive explicit and systematic instruction in the alphabetic principle and frequent opportunities for text-based reading, they learn at a faster rate than children getting standard school-based treatments. During the follow-up year, however, when children in both groups received only the standard services traditionally available in schools, rates of growth between the two groups no longer differed. Because school-based treatments (either from Chapter 1 or resource teachers) have been shown to be relatively ineffective (Kennedy et al., 1986; Moody et al., 2000; Puma et al., 1997; Snow et al., 1998), it is not completely surprising that the rate of growth of our children slowed when they returned to standard instruction. The challenge in translating research to practice is to alter standard instruction so that an accelerated growth trajectory is the norm, rather than the exception, for low-achieving children.

We were also interested in differences in response to treatment of our relatively high-skilled children (those who started above the 15th percentile on the Woodcock Basic Skills Cluster; $n = 19$, with an average score at the 19th percentile) and those who started at or below the 15th percentile ($n = 18$, with an average score at the 7th percentile). Although both groups made relatively equal gains on our standardized measures during the treatment (i.e., there were no reliable differences in gain scores between the groups), our treatment failed to close the gap between the two groups. This suggests to us that our lower skilled children might need to participate in a longer program, a more intense program (more minutes per day devoted to explicit skill-based instruction and text-based reading), and/or a program that was reinforced in the regular classroom by teachers who had also received the type of specialized training received by our tutors. The need for a longer or more intense program was reinforced by the finding (based on a review of tutor lesson plans) that lower skilled children needed to complete more lessons devoted to lower level skills (e.g., learning to read simple closed syllables, such as *sun*) than higher skilled children. Consequently, lower skilled children were less likely than the higher skilled children to complete the program and have exposure to all six syllable patterns in English. Because both initial word reading scores and level of program reached by the child at the end of the year were significantly related to end-of-year reading and spelling scores, a longer or more intense program may be especially important for the lowest achieving children.

Although the tutoring study reported here reinforces the value of an intervention that emphasizes the phonologic and orthographic connections between words, while also emphasizing fluency and text reading practice, it is important to consider other variables that might have had an influence on treatment outcomes. Several potential sources of uncontrolled differences were explored. First, our tutors received extensive training, whereas the control teachers did not. In addition, all of our tutors were certified in New York State in either special education or reading, and 10 of 12 had master's degrees. It should be noted, however, that this level of credentialing is required in New York State. That is, New York requires a master's degree for permanent certification. All of the control children who received special help in reading also received their instruction from teachers who were certified in New York State in either special education or reading, and each of these teachers had a master's degree. Thus, although we provided extensive training for our tutors and this remains a potential source of uncontrolled variance, our tutors were not exceptionally qualified in terms of certification and degree status by New York state standards, and none had used this intervention protocol before this project. We also found no evidence that some tutors were more effective than others in terms of reading and spelling outcomes.

In addition, treatment children received individual tutoring, whereas the instruction provided to the control children occurred in small groups (with an average teacher–student ratio of 1:4) and for fewer hours than the number of hours provided to the treatment children. Although we can't rule out the possibility that service delivery model (e.g., tutoring) or number of hours of instruction had an influence on reading and spelling outcomes, several sources of evidence suggest that content of instruction is more important than either service delivery model or hours of instruction. Specifically, within the control group, there was no relationship between the size of the instructional group and outcomes in either reading or spelling. This was true even after we controlled for initial skill level on the Woodcock Basic Skills Cluster. On the other hand, the literature has shown consistently that small group remedial instruction is effective across grades when the content of that instruction has a strong phonetic base (Rashotte et al., 2001), and an earlier version of this tutoring model conducted with groups of children (ranging in size from 6 to 9 children) in first- and second-grade classrooms significantly improved reading outcomes (Blachman et al., 1999). Results from a recent meta-analysis found no differences in effect sizes between phonics instruction presented one-on-one or in small groups (NRP, 2000). Thus, the literature suggests that small group instruction can be as effective as tutorial instruction when the content of that instruction reflects the content that research has shown makes a difference (Hiebert & Taylor, 2000; NRP, 2000; Snow et al., 1998). Exploring the benefits of a one-to-one intervention compared with small group treatment, while holding the content constant, would provide more empirical evidence for this claim. This question was explored recently by Vaughn and Linan-Thompson (2003), who found that struggling second-grade readers attending Title I schools made similar gains in reading when the content of instruction was held constant, whether they were tutored or whether they were instructed in small groups (1:3 ratio). Both the tutored children and the small groups, however, outperformed children receiving the same instruction in larger groups with a ratio of 1:10.

It is also the case that the treatment group received more hours of remedial reading instruction during the treatment year (from October through May) than was provided to the control children by the schools during the same period—105 hr, on average, of instruction (with a range of 86–115 hr) for treatment children and an average of 77 hr of instruction (with a range of 29–212 hr) for control children. Within the control group, however, those who received a greater number of hours of instruction did not perform better on reading and spelling measures than those who received fewer hours (or sessions) of instruction. Again, this was true even after we controlled for initial skill level on the Woodcock Basic Skills Cluster. Thus, there is no evidence that increasing the number of hours for control children—in the absence of changing the content—would have produced greater gains. The data for our control children are consistent with previous findings in the literature, in that the control children failed to gain ground, relative to their peers, on most standardized measures over the 2 years they were monitored for this study (Moody et al., 2000; Puma et al., 1997).

An interesting point, raised by one of our anonymous reviewers, provides one possible explanation for the limited effectiveness of the school-based remediation provided to the control children. As noted in the Method section, we found during screening that 52% of the children referred by the teachers were considered to be reading too well to be included in our study. Specifically, these children had a standard score on the Basic Skills Cluster of the Woodcock of 90 or above. This reviewer pointed out that if these or other similarly skilled children were participating in the school-based remediation groups that included our control children (all of whom had a standard score below 90), it is possible that the instruction might have been aimed at a level that was too high to meet the instructional needs of our control children. This is a plausible explanation, and one that should be investigated in future research.

Although both group size and number of hours of instruction are variables that could contribute to group differences, we did not find evidence that outcomes for controls varied as a function of these factors in this study. It is possible that in another replication of this study these variables might emerge as significant, but they do not appear to explain the differences between treatment and control children in this study. It would be important in future research to control for these variables by holding constant both group size and number of hours of instruction when comparing different reading protocols, and also to explore the effectiveness of differences in group size and/or number of hours of instruction while holding the instructional content constant (see Vaughn & Linan-Thompson, 2003). As pointed out by another of our anonymous reviewers, future research should also include a "trained control condition with a different [reading] focus but similarly intensive support," including equating groups on hours of teacher training, a variable not controlled in the current study.

Despite these caveats, this study adds to a growing literature (see also Rashotte et al., 2001; Torgesen et al., 2001) which demonstrates that second- and third-grade children who fail to meet the first-grade challenge of learning to read can benefit considerably in both reading and spelling from remediation that is explicit, systematic, and focused on both word-level skills and frequent opportunities for text-based reading. The content of our instruction is consistent with the research-based recommendations of recent consensus panels (NRP, 2000; Snow et al., 1998), yet there is little evidence that these instructional recommendations have been routinely incorporated into school-based remedial reading programs (RAND Reading Study Group, 2000; Shavelson & Towne, 2002). Although future research must continue to address many unanswered questions (e.g., What instructional components are most likely to strengthen comprehension for struggling readers? Will an increase in the intensity and duration of the intervention improve outcomes for the lowest-achieving children?), research-based practices can significantly improve the reading and spelling outcomes of children in remedial programs.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.

Birman, B. F., Orland, M. E., Jung, R. K., Anson, R. J., Garcia, G. N., Moore, M. T., et al. (1987). *The current operation of the Chapter 1 program*. Washington, DC: U.S. Government Printing Office.

Blachman, B. A. (1987). An alternative classroom reading program for learning disabled and other low achieving children. In R. Bowler (Ed.), *Intimacy with language: A forgotten basic in teacher education* (pp. 49–55). Baltimore: Orton Dyslexia Society.

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: Erlbaum.

Blachman, B. A., Tangel, D. M., Ball, E. W., Black, R., & McGraw, C. K. (1999). Developing phonological awareness and word recognition skills: A two-year intervention with low-income, inner-city children. *Reading and Writing: An Interdisciplinary Journal, 11,* 239–273.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Byrne, B., & Fielding-Barnsley, R. (1995). Evaluation of a program to teach phonemic awareness to young children: A 2- and 3-year follow-up and a new preschool trial. *Journal of Educational Psychology, 87,* 488–503.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33,* 934–945.

Denckla, M. B., & Rudel, R. G. (1976). Rapid "automatized" naming (R. A. N.): Dyslexia differentiated from other learning disabilities. *Neuropsychologia, 14,* 471–479.

Dole, J. A., Duffy, G. G., Roehler, L. R., & Pearson, P. D. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research, 61,* 239–264.

Ehri, L. C. (1991). Learning to read and spell words. In L. Rieben & C. A. Perfetti (Eds.), *Learning to read: Basic research and its implications* (pp. 57–73). Hillsdale, NJ: Erlbaum.

Ehri, L. C. (1997). Sight word learning in normal readers and dyslexics. In B. A. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 163–189). Mahwah, NJ: Erlbaum.

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71,* 393–447.

Elbaum, 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.

Foorman, B. R., Francis, D. J., Fletcher, J. M., & Lynn, A. (1996). Relation of phonological and orthographic processing to early reading: Comparing two approaches to regression-based, reading-level-match designs. *Journal of Educational Psychology, 88,* 639–652.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90,* 37–55.

Francis, D. J., Schatschneider, C., & Carlson, C. D. (2000). Introduction to individual growth curve analysis. In D. Drotar (Ed.), *Handbook of research in pediatric and clinical child psychology: Practical strategies and methods* (Vol. 8, pp. 51–73). New York: Plenum Press.

Fry, E. B., Kress, J. E., & Fountoukidis, D. L. (1993). *The reading teacher's book of lists* (3rd ed.). Paramus, NJ: Prentice Hall.

Gough, P. B., & Juel, C. (1991). The first stages of word recognition. In L. Rieben & C. A. Perfetti (Eds.), *Learning to read: Basic research and its implications* (pp. 47–56). Hillsdale, NJ: Erlbaum.

Hiebert, E. H., & Taylor, B. M. (2000). Beginning reading instruction: Research on early interventions. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 455–482). Mahwah, NJ: Erlbaum.

Holland, B. S., & Copenhaver, M. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin, 104,* 145–149.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

Jaccard, J., & Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression.* Thousand Oaks, CA: Sage Publications.

Kennedy, M. M., Birman, B. F., & Demaline, R. E. (1986). *The effectiveness of Chapter 1 services: An interim report from the national assessment of Chapter 1*. Washington, DC: U.S. Department of Education.

Levy, B. A. (2001). Moving the bottom: Improving reading fluency. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 357–379). Timonium, MD: York Press.

Liberman, I. Y., & Liberman, A. M. (1990). Whole language vs. code emphasis: Underlying assumptions and their implications for reading instruction. *Annals of Dyslexia, 40,* 51–76.

Lyon, G. R. (2001, March 8). *Measuring success: Using assessments and accountability to raise student achievement (Statement to the Committee on Education and the Workforce)*. Retrieved July 28, 2002, from http://edworkforce.house.gov/hearings/107th/edr/account3801/lyon.htm

Lyon, G. R., & Fletcher, J. M. (2001, Summer). Early warning systems. *Education Matters, 1*(2), 22–29. Retrieved July 28, 2002 from http://www.educationnext.org/20012/22.html

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., et al. (2001). Rethinking learning disabilities. In C. E. Finn, Jr., R. A. J. Rotherham, & C. R. Hokanson, Jr. (Eds.), *Rethinking special education for a new century* (pp. 259–287). Washington, DC: Thomas B. Fordham Foundation and Progressive Policy Institute.

Makar, B. W. (1995). *Primary phonics.* Cambridge, MA: Educators Publishing Service.

Maxwell, S. E. (1998). Longitudinal designs in randomized group comparisons: When will intermediate observations increase statistical power? *Psychological Methods, 3,* 275–290.

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.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implication for reading instruction: Reports of the subgroups*. Washington, DC: National Institute of Child Health and Human Development.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Pressley, M. (2002). *Reading instruction that works: The case for balanced teaching* (2nd ed.). New York: Guilford Press.

Puma, M. J., Karweit, N., Price, C., Ricciuti, A., Thompson, W., & Vaden-Kiernan, M. (1997). *Prospects: Final report on student outcomes*. Cambridge, MA: ABT Associates. (ERIC Document Reproduction Service No. ED413411)

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Rashotte, C. A., MacPhee, K., & Torgesen, J. K. (2001). The effectiveness of a group reading instruction program with poor readers in multiple grades. *Learning Disability Quarterly, 24,* 119–134.

SAS Institute. (2001). SAS OnlineDoc (Version 8.1) [Computer software]. Cary, NC: Author.

Seaman, M. A., Levin, K. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110,* 577–586.

Share, D. L., & Stanovich, K. E. (1995). Cognitive processes in early reading development: Accommodating individual differences into a model of acquisition. *Issues in Education, 1,* 1–57.

Shavelson, R. J., & Towne, R. (Eds.). (2002). *Scientific research in education*. Washington, DC: National Academy Press.

Shaywitz, B. A., Shaywitz, S. E., Blachman, B. A., Pugh, K. R., Fulbright, R. K., Skudlarski, P., et al. (2004). Development of left occipitotemporal systems for skilled reading in children after a phonologically-based intervention. *Biological Psychiatry, 55,* 926–933.

Shaywitz, S. E., & Shaywitz, B. A. (1996). Unlocking learning disabilities: The neurological basis. In S. C. Cramer & W. Ellis (Eds.), *Learning disabilities: Lifelong issues* (pp. 255–260). Baltimore: Paul H. Brookes.

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

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360–407.

Stuebing, K. K., Fletcher, J. M., LeDoux, J. M., Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2002). Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal, 39,* 469–518.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice, 15,* 55–64.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34,* 33–58, 78.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency (TOWRE)*. Austin, TX: Pro-ed.

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.

Uitenbroek, D. G. (1997). *SISA binomial*. Southampton, England: Author. Retrieved January 1, 2004, from http://home.clara.net/sisa/binomial.htm

Vaughn, S., & Linan-Thompson, S. (2003). Group size and time allotted to intervention: Effects for students with reading difficulties. In B. R. Foorman (Ed.), *Preventing and remediating reading difficulties: Bringing science to scale* (pp. 299–324). Baltimore: York Press.

Vellutino, F. R. (1991). Introduction to three studies on reading acquisition: Convergent findings on theoretical foundations of code-oriented versus whole-language approaches to reading instruction. *Journal of Educational Psychology, 83,* 437–443.

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.

Vellutino, F. R., Scanlon, D. M., & Tanzman, M. S. (1994). Components of reading ability: Issues and problems in operationalizing word identification, phonological coding, and orthographic coding. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 279–332). Baltimore: Paul Brookes.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processes (CTOPP).* Austin, TX: Pro-ed.

Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher, 32*(7), 22–30.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third Edition (WISC–III).* San Antonio, TX: Psychological Corporation.

Wiederholt, J. L., & Bryant, B. R. (1992). *Gray Oral Reading Tests—Third Edition (GORT–3).* Austin, TX: Pro-ed.

Wilkinson, G. S. (1993). *The Wide Range Achievement Test 3 (WRAT3).* Wilmington, DE: Wide Range.

Williams, J. P. (1994). Twenty years of research on reading: Answers and questions. In F. Lehr & J. Osborn (Eds.), *Reading, language, and literacy: Instruction for the twenty-first century* (pp. 59–73). Hillsdale, NJ: Erlbaum.

Williams, J. P., Lauer, K. D., Hall, K. M., Lord, K. M., Gugga, S. S., Bak, S., et al. (2002). Teaching elementary school students to identify story themes. *Journal of Educational Psychology, 94,* 235–248.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests—Revised (WRMT–R).* Circle Pines, MN: American Guidance Service.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery—Revised (WJ–R).* Chicago: Riverside.

Woodcock, R. W., & Mather, N. (1989). WJ–R Tests of Achievement: Examiner's manual. In R. W. Woodcock & M. B. Johnson, *Woodcock-Johnson Psycho-Educational Battery—Revised (WJ–R).* Chicago: Riverside.

Received April 8, 2003
Revision received February 16, 2004
Accepted April 5, 2004 ■