

A Randomized Experimental Evaluation of the Impact of Accelerated Reader/Reading Renaissance Implementation on Reading Achievement in Grades 3 to 6

John A. Nunnery

*Department of Educational Leadership and Counseling
Old Dominion University*

Steven M. Ross and Aaron McDonald

*Center for Research in Educational Policy
University of Memphis*

Components of the School Renaissance® program, including Accelerated Reader and Reading Renaissance, have been implemented in more than 65,000 schools in the United States. Despite the program's popularity, there have been no published, well-controlled evaluations of its effectiveness. This randomized field experiment was designed to gauge program impacts on the reading achievement of 978 urban students in Grades 3 to 6. Schools and teachers within schools volunteered to participate in the study, with the foreknowledge that teachers would be randomly assigned to either implement School Renaissance or serve as controls. A 3-level hierarchical linear model was used to estimate the impact of the program on student reading growth trajectories on the STAR Reading test. Students in School Renaissance classrooms exhibited significantly higher growth rates than those in control classrooms, with effect size estimates ranging from +0.07 to +0.34 across grades. Quality of program implementation did not predict student achievement growth but was correlated with a reduction in the negative effect of learning disability status.

The School Renaissance program, particularly the Accelerated Reader (AR) and Reading Renaissance (RR) components, is one of the most widely implemented

school reform programs in the United States. In 2004, components of School Renaissance were being implemented in more than 65,000 schools in the United States (National Clearinghouse on Comprehensive School Reform [NCCSR], 2004). Although AR is viewed primarily as a supplementary reading program, implementation of AR in conjunction with RR meets the criteria for designation as a Comprehensive School Reform (CSR) model (NCCSR, 2004). This designation permits schools to apply for CSR grants totaling as much as \$225,000 over 3 years to implement the model. Actual costs of implementation, which vary depending on school size and components implemented, range from \$30,000 to \$75,000 per school per year (NCCSR, 2004). As of January 2004, 250 schools had received CSR grants to implement RR and AR; of these, 61 had also received support to implement math components (NCCSR, 2004).

AR is a computerized information system that provides students and teachers with immediate diagnostic feedback on student reading practice through short quizzes. AR facilitates guided reading practice by using feedback from AR quizzes to help students and teachers select books at the appropriate level, monitor comprehension of books read, and guide further reading practice. AR tracks three types of reading practice (reading aloud, paired reading, and independent reading) for a variety of texts including student self-selected books, textbooks, and magazines. In addition, teachers can manipulate the reports provided by AR to diagnose reading difficulties for individual students or groupings of students.

AR is the core component of an approach to guided reading practice called RR. RR is a set of practices designed around six key principles. The first principle is that students need sufficient opportunities to practice reading to become better at it (Anderson, Hiebert, Scott, & Wilkinson, 1985; Snow, Burns, & Griffin, 1998). RR recommends 30 to 60 min of reading time. The second principle is that the benefit of reading practice is optimized when that practice is at a high level of success, which is defined as 85% correct and above on AR quizzes (Paul, 2003). Third, students should read books matched to their abilities, defined as their zone of proximal development. Practice that is too easy does little to improve skills, and practice that is too difficult leads to frustration (Allington, 2001; Anderson et al., 1985; Betts, 1946; Brophy & Good, 1986; Snow et al., 1998). In between lies a student's zone of proximal development (Vygotsky, 1978), a level of difficulty that leads to optimal learning. The fourth RR principle involves information feedback through the following means. AR provides daily and weekly feedback on student comprehension of books they have read; teachers conduct a daily "status of the class" that involves monitoring and conferencing with each student on a daily basis; and at a minimum of three times per year, students take computer-adaptive assessments such as STAR Reading or STAR Early Literacy. This kind of feedback can help students be more attentive and careful when they are doing their work (Black & William, 1998; Samuels & Wu, 2003; Walberg, 1984). The fifth principle is the establishment of each student's personal goals for reading practice time, book level,

and performance on quizzes. The establishment of goals has been associated with strong effects on student learning (Lipsey & Wilson, 1993; Marzano, 2003; Walberg, 1999; Wise & Okey, 1983). The sixth principle is that teachers provide personalized instruction based on the information provided by AR, periodic assessments, and daily monitoring of student progress. Meta-analyses conducted by Walberg (1984) and Bloom (1984) have confirmed the effectiveness of adjusting instruction to address the individual needs of students.

In practice, a primary component of RR is a significant daily block of time (30–60 min) devoted to reading practice within the To-With-Independent (TWI) framework. The TWI framework divides the time into three segments: reading texts to a child (T), reading texts with the child using a paired-reading technique (W), and allowing the child to read independently (I). Texts are normally self-selected trade books, but they can also include assigned basal texts or magazine articles. In pre-K through Grade 3, the reading practice is heavily weighted toward the T and W elements. As children develop decoding skills, they transition into increasing amounts of independent reading. In all cases, student ability level is taken into account as teachers guide students to texts that they can enjoy and comprehend at a high level.

On completion of reading a text, whether in T, W, or I mode, each student takes a short, literal comprehension quiz on the computer using AR. The quiz is immediately scored, and AR generates a report for the student that indicates his or her performance. This report facilitates a discussion with the teacher, who provides reinforcement or correction as needed. Quiz performance data are stored in the AR database and are frequently accessed by teachers to monitor the success rates of students. Teachers can easily identify the students who are experiencing successful reading practice and those who are not. With this information, the teacher can intervene as needed, whether by more closely monitoring the student's book selection habits, or providing targeted instruction on a particular reading skill with which the student may be struggling.

RESEARCH ON THE EFFECTIVENESS OF AR/RR

Despite the apparent widespread popularity of the program and "evidence of effectiveness" criterion employed by reviewers to determine CSR eligibility, there are few published, controlled studies of the effects of AR/RR on student achievement in reading. Critics of the program claim that the program's use of incentives reduces students' intrinsic motivation to read, and that the program constricts each student's choice of reading materials (Biggers, 2001; Carter, 1996; Stevenson & Camarata, 2000). They also point to fact that the preponderance of evidence of the program's effectiveness is based on correlational data, anecdotal reports of increased tests scores from sites implementing the program, or gains of program stu-

dents compared to norm gains from pre-experimental (i.e., no comparison or control group) designs (Krashen, 2003).

Indeed, numerous pre-experimental studies and anecdotal reports have documented gains in reading achievement for AR/RR students but lack a control group to which one can compare results (Arkebauer, MacDonald, & Palmer, 2002; Cuddeback & Ceprano, 2002). School- or district-wide improvements in reading achievement after implementation of AR or AR/RR have been noted in reports from locations in states across the country; at all grade levels; in schools serving impoverished and middle-class student populations; in schools serving limited English proficient (LEP) students; and in rural, suburban, and urban areas (e.g., Cutler, 2002; Fine, 2001; Howard, 1999; Morris, 2001).

Large-scale correlational studies also point to a relation between AR/RR implementation and improvement in student reading scores on standardized tests. In a study involving more than 60,000 students in Tennessee, Topping and Sanders (2000) found that teacher value-added effectiveness scores in reading were correlated with the amount of reading and average quiz scores in AR classrooms, and that teachers who received RR training had higher effectiveness scores than those who did not. Terrance, VanderZee, Rue, and Swanson (1996) examined the relation between school-level performance and implementation of AR in 2,500 AR schools and 3,500 comparison schools, finding that AR schools had significantly higher achievement gains than demographically similar comparison schools, particularly in urban schools serving socioeconomically disadvantaged students. Two small-scale quasi-experimental studies reinforce the findings that AR/RR is particularly effective with at-risk (low socioeconomic status) students and students with learning disabilities (Kamarian, 2001; Scott, 1999).

Other quasi-experimental studies indicate that even poor implementations of AR/RR may have a positive effect on student achievement. Vollands, Topping, and Evans (1999) found that AR implementation resulted in greater gains in reading among at-risk third graders than those achieved by comparison teachers implementing an alternative program, even though the AR implementation was not particularly strong. Holmes and Brown (2002) conducted a quasi-experimental study of AR/RR effects in two Georgia elementary schools, one with a high-poverty, predominantly African American student population, the other a moderate-poverty (50% eligible for free lunch), rural, and predominantly White student population. In analyses of covariance (ANCOVAs) comparing these schools to matched control schools, Iowa Test of Basic Skills scores were used as a covariate, and Georgia Criterion-Referenced Competency Test (GCRCT) scores and STAR Reading assessments served as outcomes. Across all grades, the average effect size reported for reading was +0.50 on GCRCT measures and +0.09 on STAR measures. Program effects were significant on all GCRCT outcomes and on STAR Reading assessments in Grades 3 and 4.

The few controlled longitudinal studies of AR/RR provide a mixed view of its effectiveness. Sadusky and Brem (2002) performed an ex post facto analysis of reading achievement of students in a school implementing AR/RR and those in a demographically matched comparison school that used only AR. The authors reported a statistically significant increase of 9 points on Stanford Achievement Test, Version 9 (SAT-9) scores in the AR/RR school over a 6-year period, compared to a nonsignificant increase of 2 points on the SAT-9 in the AR-only school. Using a similar design, Peak and Dewalt (1993) found that middle school students enrolled at an AR school had higher mean reading on California Achievement Test (CAT) reading scores than those enrolled in a comparison school after using the program for several years. In contrast to the positive findings of these studies, Pavonetti, Brimmer, and Cipielewski (2000) reported that comparison students were doing more reading in middle school than students who participated in AR in elementary school as measured on a title recognition instrument.

Two small-scale quasi-experimental studies also report mixed results of the program, although both studies suffer from a confounding of teacher and program, and were conducted over relatively short time frames (Facemire, 2000; Toro, 2001). The only true experimental evaluation of AR effects, in which AR implementation was randomly assigned to teachers, was conducted by Samuels, Lewis, Wu, Reininger, and Murphy (2003). Samuels et al. reported that students in the AR condition scored significantly higher in reading comprehension on the GRADE reading test after 10 weeks, achieving approximately three times the gain of control students. This study was conducted within a single school, however, so the possibility for confounding teacher effects and program effects was high. The authors also did not provide mean scores or the results of statistical tests, making the study difficult to evaluate.

The existence of many positive anecdotal reports and uncontrolled studies suggests that AR/RR effects may be generalizable across many contexts and populations but provides limited robust evidence that the program actually improves reading performance above what might otherwise be expected. Likewise, the large-scale correlational studies, although providing tantalizing suggestive evidence, lack the control necessary to test for a causal relation between AR/RR implementation and student achievement. Because implementation is measured to a large extent by how well students perform on STAR Reading tests and the number of books read, the argument from implementation to higher reading achievement seems somewhat tautological. Evidence of AR/RR effectiveness from experimental and quasi-experimental studies is rather weak, given the presence of confounding variables and the short time frames over which the studies were conducted. Indeed, evidence of the effectiveness of any CSR models from randomized control group designs is rare. In a recent meta-analysis of CSR effects, Borman, Hewes, Overman, and Brown (2003) found that only 3% of all studies of CSR effects used such designs.

PURPOSES OF THIS STUDY

Given the enormous popularity of AR/RR, past criticisms of the program, and the relatively weak evidence supporting its effectiveness, the major purpose of this study was to conduct an experimental evaluation of AR/RR effects on the reading achievement of students attending urban, high-poverty elementary schools. Secondary goals included examining the relation between AR/RR implementation level and student achievement, and, following Scott's (1999) findings, to explore the relation between AR/RR implementation and the reading achievement of students with learning disabilities. To these ends, this study addressed the following research questions:

- What impact does implementation of AR/RR have on growth in reading achievement for at-risk students in Grades 3 to 6?
- How does the quality of AR/RR implementation relate to student growth in reading?
- Does the quality of AR/RR implementation have an effect on growth in reading for students with learning disabilities?

METHOD

Participants

Study participants included 978 students in Grades 3 ($n = 250$), 4 ($n = 381$), 5 ($n = 215$), and 6 ($n = 132$), and 44 teachers in a large urban school district in the southern United States. Of the students, 89.9% were African American, 83% were eligible for free or reduced-price lunch, and 3.3% had a specified learning disability. Slightly more than half (53.5%) of students were female. Study participants attended or worked in one of nine inner-city schools in a large school district in the southern United States.

Program Description

Teachers assigned to the treatment condition simultaneously implemented AR and RR. AR is a computerized curriculum management program in which students choose books to read and complete a brief, multiple-choice comprehension quiz. A 60% correct score is required to pass an individual quiz, and an 85% quiz average is considered a mastery level across books of similar difficulty. Based on quiz results, the program generates a list of selected readings appropriate to the student's reading level. RR is a teacher professional development program designed to facilitate teachers' use of several practices, including providing 60 min per day for stu-

dent reading, use of AR in the classroom, managing students' use of reading logs, identifying students' zones of proximal development to identify appropriately challenging reading materials, and use of AR diagnostic reports to identify students who need remediation or other interventions. Unlike some previous implementations of AR, this study involved no extrinsic incentives.

Control group curriculum. The participating district required a 90-min reading block in the participating grades. All elementary schools in the study used the same commercially available basal reading program. A suggested schedule for small- and whole-group activities, by grade level, was provided for K–6 teachers. Participating schools were implementing sustained silent reading programs to support fluency, comprehension, and vocabulary development. The district also had a stated goal of 25 books read per student.

Student Achievement Measure

The STAR Reading test was administered to participating students in the control and treatment conditions. This computerized test uses Rasch measurement techniques to generate a scaled score ranging from 0 to 1400, which ranges across grade levels. In a 1999 norming study, split-half reliability coefficients for STAR Reading ranged between .89 and .90 for third through sixth grades (STAR Reading, 2001). STAR Reading scale scores also exhibit a moderate to strong correlation with other standardized reading tests, including the California Achievement Test, the Comprehensive Test of Basic Skills, the Gates-MacGinitie Reading Test, the Stanford Achievement Test, and the Iowa Test of Basic Skills, with validity coefficients ranging from .36 to .97 in Grades 1 to 6 (Nebelsick-Gullett, 2003). Sadusky and Brem (2002) reported that correlation coefficients between SAT–9 and the STAR Reading test scores ranged between +.65 and +.75 across the elementary grades. The STAR Reading test was administered in September, January, and April to provide a longitudinal profile of reading achievement.

Program Implementation Measures

Implementation ratings were generated by consultants based on three on-site consultations during the school year. The consultant ratings were gathered in four areas: (a) classroom implementation, which rated the quality of implementation of the AR/RR program for each classroom on a 3-point Likert-type scale ranging from 1 (*poor*) to 3 (*strong*); (b) TWI (time spent reading to, with, or independently) rated on a 3-point Likert-type scale ranging from 1 (*30 min or less*) to 3 (*60 min or more*); (c) hardware usage, which represented the degree to which problems were experienced using the computer hardware and software necessary to implement the program, with scores ranging from 1 (*no problems*) to 3 (*substantial problems*);

and (d) principal support, which rated the perceived level of principal support for the program on a 3-point Likert-type scale ranging from 1 (*poor*) to 3 (*strong*). Additionally, distance consulting implementation measures, which are intended to provide feedback for program monitoring, were collected for each teacher in the treatment group. These measures included (a) the percentage of students reading at or above the expected level, (b) reading practice points achieved, and (c) a measure of independent reading practice. To facilitate further analyses, a principal components analysis was performed to generate a single, regression-based factor score representing overall classroom-level implementation. A single factor accounted for 79.4% of the variance in the seven implementation measures.

Procedure

Elementary schools within the participating district were notified of the opportunity to participate in the study. Schools that had principal consent and at least two teacher volunteers at any grade level were eligible to participate in the study. Teachers volunteered with the understanding that they would be randomly assigned to either a treatment group, meaning they would implement AR and RR in their classrooms, or a control group that would not implement either program or participate in professional development related to either program. After the pool of participating teachers was determined, teachers were randomly assigned to pairs within grade levels. Within each pair, one teacher was then randomly assigned to the control condition. In cases with odd numbers of teacher volunteers within a grade level, a set of three teachers was randomly determined, with one teacher of the set randomly assigned to the control condition. Teachers in the control condition were promised the opportunity, if desired, to implement AR/RR in their classrooms the following year.

Each school received on-site consulting once a month from two Renaissance consultants. During a normal visit, one consultant met with administration, and the second consultant met with teachers to troubleshoot technical issues and provide implementation feedback. Each consultant kept an implementation log and at the end of the year completed implementation ratings based on their observations and interviews. AR and control students were administered the STAR tests during the same time periods in September (pretest), January (midterm), and April (posttest). The pretest STAR test was used by AR teachers to gauge students' initial reading levels. There was no further use of STAR as part of the program, and treatment and control students took the STAR test an equal number of times.

Initial Data Screening and Transformations

Normality and variance screening. STAR scale scores at pretest, midterm, and posttest were first examined for normality within grade levels and homo-

geneity of variance across grade levels. Initial examination of the test scores revealed moderate to strong skewness for pretest ($Sk = 0.72$), midterm ($Sk = 1.10$), and posttest scores ($Sk = 1.22$). Levene's test indicated that variances across grade levels were significantly different for pretest, $F(3, 877) = 9.77, p < .001$; midterm, $F(3, 798) = 13.21, p < .001$; and posttest, $F(3, 828) = 6.92, p < .01$. A square root transformation was performed on all three variables to reduce skewness and stabilize variances. Levene's test indicated no significant differences in variance across grade levels on the transformed variable. Skewness was also substantially reduced through transformation for each variable: pretest ($Sk = -0.02$), midterm ($Sk = 0.14$), and posttest ($Sk = 0.24$). Visual examination of normal plots within grade levels and spread-and-level plots confirmed that the square root transformation was largely successful in inducing normality and homogeneity of variance.

A total of 1,023 students in Grades 3 to 6 were enrolled in either AR classrooms or control classrooms for the entirety of the study. Of these, 4.4% ($n = 45$) did not complete any STAR Reading tests, and thus were eliminated from subsequent analyses. Chi-square tests indicated no relation between missing test data and treatment condition, free or reduced-price lunch status of student, learning disability status of student, or gender of student.

Multiple imputation of missing values. Initial screening procedures showed that, of 978 students who had data from at least one test administration, 97 were missing pretest scores, 176 were missing midterm scores, and 146 were missing posttest scores. About 75% had matching pretest-midterm scores ($n = 736$), matching pretest-posttest scores ($n = 742$), or matching midterm-posttest scores ($n = 738$). About 70% ($n = 679$) had scores for all three test administrations, whereas 11% ($n = 113$) had only one test score. Chi-square tests indicated no significant relations between treatment or grade level with the pattern of missing data. Due to the large percentages of students who had complete data for at least two scores, it was deemed desirable to use multiple imputation procedures so all cases could be included in analyses of program effects. Multiple imputation uses expected maximum likelihood methods to generate values to replace missing data. The imputed values represent the most likely value that would have been observed for a particular case given the data profile on measured variables.

Analysis. A three-level hierarchical linear model was estimated to examine relations between student characteristics and growth in reading achievement, and to estimate the impact of treatment condition on the growth trajectory of students. The Level 1 model was a within-students model relating the repeated measures of reading achievement to time, where Time 0 was the September administration, Time 1 was the January administration, and Time 2 was the April administration:

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{time}) + e_{ijk} \quad (\text{within-students model})$$

In the Level 2 (within-classroom) model, the Level 1 intercept (π_{0jk}) and Level 1 slope (π_{1jk}) were modeled as functions of student characteristics. Specifically, the intercept was modeled as a function of grade level and gender, whereas the slope was modeled as a function of the learning disability status of the child:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{grade}) + \beta_{02k}(\text{gender}) + r_{0jk} \quad (\text{within-classes intercept model})$$

$$\pi_{1jk} = \beta_{10k} + \beta_{11k}(\text{learning disability}) + r_{1jk} \quad (\text{within-classes slope model}).$$

Although each of the β s is a potential outcome in the Level 3 model, the interest in this study was to explain variation in β_{10k} , the average time slope within each classroom k . Thus, the Level 3 (between-classrooms) model was:

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + u_{00}; \\ \beta_{01k} &= \gamma_{010}; \\ \beta_{02k} &= \gamma_{020} + u_{02}; \\ \beta_{10k} &= \gamma_{100} + \gamma_{101}(\text{treatment condition}) + u_{10}; \\ \beta_{11k} &= \gamma_{110} + u_{11}. \end{aligned}$$

The residual parameter variance for grade (β_{01k}) was set to zero. A model with only the Level 1 predictor (time) was also estimated to serve as a baseline to assess model fit. Exploratory analyses were performed to test for the effects of variations in program implementation and classroom heterogeneity (i.e., classroom achievement variance at Time 0) on Level 3 slopes and intercepts.

To aid in the interpretation of results, the mean transformed STAR Reading score was computed and plotted for each administration, grade level, and treatment condition. Treatment effect size estimates were computed at each grade level by performing ANCOVA using treatment condition as the independent variable, untransformed STAR Reading scale score at Time 2 as the dependent variable, and untransformed STAR Reading scale score at Time 0 as a covariate. Cohen's d was computed as an effect size estimate for each grade level, performing an ANCOVA using treatment condition as the independent variable, STAR Reading scale score at Time 2 as a dependent variable. The adjusted mean for the control group was subtracted from the adjusted mean for the treatment group, and the result was divided by the pooled within-groups standard deviation for STAR Reading at Time 0 to yield d . Descriptive statistics of the untransformed STAR scores were computed for each grade level and treatment condition, including the standardized difference between the treatment and control group means. The standardized difference between the means was computed by subtracting the control group mean from the treatment group mean, then dividing by the con-

trol group standard deviation. Finally, frequencies of consultants' implementation ratings were computed for each grade level.

RESULTS

Descriptive Statistics

Untransformed STAR pretest mean differences between control and treatment students were quite small in third grade ($M_C = 241.8$ vs. $M_T = 249.4$) and fourth grade ($M_C = 335.6$ vs. $M_T = 333.8$; see Table 1). Standardized pretest mean differences at these grade levels were +0.07 and -0.01 for third and fourth grade, respectively (see Figure 1). Pretest mean differences in fifth and sixth grade substantially favored treatment students ($M_C = 397.4$ vs. $M_T = 436.5$ for fifth grade, $M_C = 519.8$ vs. $M_T = 564.9$ for sixth grade; see Table 1), corresponding to relatively large standardized mean differences of +0.27 and +0.24. As indicated in Figure 1, AR/RR students in third and fourth grade had progressively higher standardized mean differences from pretest through posttest (from +0.07 at pretest to +0.50 at posttest for third grade, and from -0.01 at pretest to +0.14 at posttest for fourth grade). However, AR/RR students in fifth and sixth grade showed uneven progress relative to control students across all three testing occasions, with unsubstantial relative gains overall (from +0.27 to +0.31 for fifth grade, and +0.24 to +0.27 for sixth grade; see Figure 1). Overall, the descriptive profile of test results indicated a strong program

TABLE 1
STAR Reading Scale Score Means by Grade and Treatment Status

Grade/Treatment	Pretest			Midtest			Posttest		
	M	SD	n	M	SD	n	M	SD	n
Third									
Control	241.8	116.7	112	278.7	112.7	112	315.6	134.7	112
AR/RR	249.4	121.9	138	307.5	124.9	138	383.3	200.4	138
Fourth									
Control	335.6	138.6	176	369.9	132.6	176	385.3	138.6	176
AR/RR	333.8	132.3	205	379.0	135.2	205	405.0	136.4	176
Fifth									
Control	397.4	144.9	94	412.7	146.9	94	435.1	158.1	94
AR/RR	436.5	176.6	121	472.3	186.3	121	484.6	183.2	121
Sixth									
Control	519.8	189.8	59	596.5	218.8	59	592.3	221.4	59
AR/RR	564.9	178.7	73	614.8	207.2	73	652.9	220.8	73

Note. AR/RR = Accelerated Reader/Reading Renaissance.

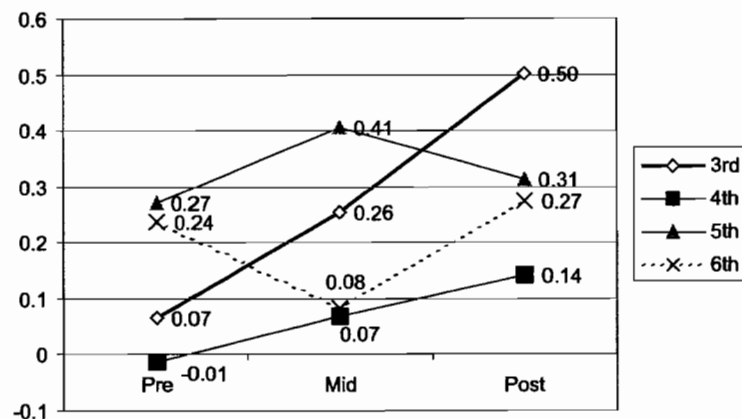


FIGURE 1 Standardized mean difference between treatment and control groups by grade level and test administration: Untransformed STAR Reading scale scores.

effect in third grade, a small to moderate effect in fourth grade, and no effect in fifth or sixth grade.

Unconditional Hierarchical Linear Model

The reliability estimates of the random Level 2 coefficients for β_{00} (achievement at Time 0) and for β_{10} (time slope) were .927 and .852, respectively. These values indicated a strong possibility of discriminating between classes with respect to baseline achievement and growth in reading achievement. The average time-slope coefficient was statistically significant, $\gamma_{100} = 0.948$, $t(43) = 9.56$, $p < .001$, indicating that, on average, student growth rates were positive across classrooms. A test of the Level 3 variance components indicated significant variability across classrooms in the mean rate of change in reading achievement, $\chi^2(43, N = 44) = 332.89$, $p < .001$. The correlation between the Level 1 intercepts and Level 1 slopes was $-.065$, showing little or no relation between the beginning level of achievement and growth in achievement.

Conditional Model: Treatment Status

As expected, grade level ($\gamma_{010} = 2.30$, $t = 12.74$, $p < .001$) and gender ($\gamma_{020} = 1.26$, $t = 6.20$, $p < .001$) were significantly related to STAR Reading scores at Time 0, reflecting higher pretest scores at higher grade levels and higher pretest scores for girls (see Table 1). The Level 3 intercept for time slope was significant ($\gamma_{100} = 0.76$,

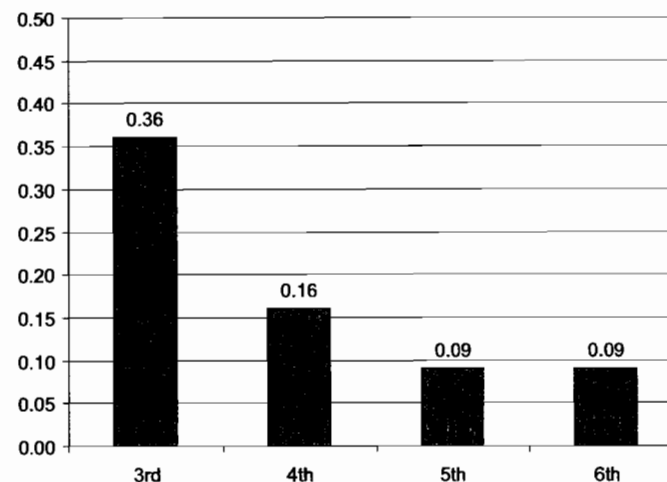


FIGURE 2 Effect size estimates by grade level.

$t = 8.41$, $p < .001$), as were the relations between the time slope with treatment ($\gamma_{101} = 0.39$, $t = 2.41$, $p = .02$) and learning disability status ($\gamma_{110} = -0.60$, $t = -2.53$, $p = .015$). Effect size estimates based on the ANCOVA previously described were $+0.36$, $+0.16$, $+0.09$, and $+0.09$ for third, fourth, fifth, and sixth grades, respectively (see Figure 2). These results parallel those of the descriptive analysis, indicating a large positive program effect in Grade 3, a small to moderate positive effect in Grade 4, and small positive effects in Grades 5 and 6.

Conditional Model: Classroom Heterogeneity and Program Implementation Effects

Correlation coefficients between classroom heterogeneity and program implementation scores with Level 3 slopes and intercepts were examined to determine whether the addition of either of these variables influenced the slope for learning disability. Based on this examination, program implementation was included as a predictor of the slope for learning disability status. Program implementation scores ($\gamma_{101} = 0.16$, $t = 1.79$, $p = .08$) did not predict time slopes as well as the simple treatment status indicator. As reported in Table 2, the level of program implementation probably has a positive relation with the average impact of learning disability status across schools ($\gamma_{111} = 0.54$, $t = 1.89$, $p = .065$), indicating that high-implementation classrooms tended to mitigate the negative effects of learning disability on student growth in reading achievement.

TABLE 2
Treatment and Learning Disability Effects on Growth
in Student Reading Achievement

Fixed Effect	Coefficient	SE ^a	<i>t</i>	<i>df</i>	<i>p</i>
Intercept					
Intercept (γ_{000})	6.78	0.95	7.13	43	.000
Grade (γ_{010})	2.30	0.18	12.74	975	.000
Gender (γ_{020})	1.26	0.20	6.20	43	.000
Time slope					
Intercept (γ_{100})	0.76	0.09	8.41	42	.000
Treatment (γ_{101})	0.39	0.16	2.41	42	.020
Learning disability (γ_{110})	-0.60	0.24	-2.53	43	.015

^aRobust standard error.

The relation between program implementation and learning disability status was further explored by computing empirical Bayes (EB) estimates of the classroom-level learning disability slopes, classifying AR classrooms as either low- or high-implementation based on a median split of the implementation factor scores, then performing a one-way analysis of variance with treatment status (control, low-implementation, high-implementation) as the independent variable and EB slope estimates as the dependent variable. The results indicated a significant effect of treatment status on learning disability slopes, $F(2, 41) = 5.55$, $p = .007$. Pairwise post hoc comparisons were performed using Fisher's least significant difference procedure, which indicated that learning disability slopes in high-implementation classrooms ($M = -0.20$) were significantly higher than slopes in either control ($M = -0.62$) or low-implementation ($M = -0.92$) classrooms; no significant difference was observed between control and low-implementation classrooms (see Figure 2).

Implementation ratings. As reported in Table 3, consultants generally rated classroom implementation as average for all grade levels except fifth grade, in which 60% ($n = 3$) of the ratings were poor. Likewise, consultant estimates of TWI were quite low for fifth grade, with 80% ($n = 4$) of fifth-grade AR/RR teachers' classrooms rated as spending less than 30 min engaged in these reading activities. Fourth- (44%) and fifth-grade (40%) teachers tended to experience more substantial problems related to using program technology. Principal support was rated as weak in 44% ($n = 4$), average in 22% ($n = 2$), and strong in 33% ($n = 3$) of program schools.

DISCUSSION

Students in AR/RR classrooms had significantly higher growth rates in reading compared to students in control classrooms. Effect size estimates were higher in

TABLE 3
Treatment and Implementation Level Effects on Growth
in Student Reading Achievement

Fixed Effect	Coefficient	SE ^a	<i>t</i>	<i>df</i>	<i>p</i>
Intercept					
Intercept (γ_{000})	6.76	0.95	7.09	43	.000
Grade (γ_{010})	2.30	0.18	12.82	975	.000
Gender (γ_{020})	1.26	0.20	6.14	43	.000
Time slope					
Intercept (γ_{100})	0.76	0.09	8.41	42	.000
Treatment (γ_{101})	0.39	0.16	2.39	42	.022
Learning disability					
Intercept (γ_{110})	-0.59	0.23	-2.50	42	.017
Implementation (γ_{111})	0.54	0.28	1.89	42	.065

^aRobust standard error.

lower grade levels: +0.36 in third grade, +0.16 in fourth grade, +0.09 in fifth grade, and +0.09 in sixth grade. The effect size estimates reported in this study are commensurate with median effect size estimates reported from comparison group studies for CSR models with the "strongest evidence of effectiveness," which were $d = +0.15$ for Direct Instruction, $d = +0.05$ for the School Development Program, and $d = +0.18$ for success for all (Borman et al., 2003).

Like the effect size estimates, implementation ratings also tended to be higher in lower grades. Nevertheless, accounting for variation in implementation did not improve the fit of the model over simply knowing whether the teacher was in the AR/RR treatment condition or the control condition. As found in previous controlled studies of AR/RR, the program produced significant achievement effects in spite of less than optimal implementation (Holmes & Brown, 2002; Vollands et al., 1999). Compared to many of the CSR models, such as those developed as part of New American Schools (e.g., see Berends, Chun, Schuyler, Stockley, & Briggs, 2002), implementation of AR/RR is fairly straightforward and does not involve complex curricula or structural changes in the classroom. Thus, loss of some implementation fidelity may be less damaging to program effects than with more complex or comprehensive whole-school reforms. It is also possible that the effects observed were produced by treatment students reading more books than control students as a result of the AR component of the program—the implementation measure predominantly reflected implementation of the classroom instruction component.

Exploratory analyses suggested a positive relation between AR/RR implementation and reading achievement growth rates for students with designated learning disabilities. Follow-up analyses indicated that high-implementation AR/RR classrooms significantly reduced the negative impact of learning disability status on growth in reading when compared to control classrooms or low-implementation

classrooms. This finding extends those of Scott (1999), who reported positive effects of AR on the reading achievement of middle school students with learning disabilities, and of Holmes and Brown (2002), who found that AR was more effective with lower achieving students.

In summary, the implementation of AR and RR had consistently positive effects on the reading achievement of at-risk students across Grades 3 to 6, with larger effects in the earlier grades and small effects in the upper grades. Surprisingly, fidelity of program implementation at the classroom level did not predict achievement any better than simple knowledge of whether the classroom was implementing the program, although children with learning disabilities in high-implementation classrooms had significantly higher achievement gains than children with learning disabilities in either control or low-implementation classrooms.

REFERENCES

- Allington, R. L. (2001). *What really matters for struggling readers: Designing research-based programs*. New York: Longman.
- Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report on the Commission of Reading*. Washington, DC: National Institute of Education.
- Arkebauer, C., MacDonald, C., & Palmer, C. (2002). *Improving reading achievement through the implementation of a balanced literacy approach*. Master of Arts Research Project, Saint Xavier University and SkyLight Professional Development Field-Based Program. (ERIC Document Reproduction Service No. ED471063)
- Berends, M., Chun, J., Schuyler, G., Stockley, S., & Briggs, R. J. (2002). *Challenges of conflicting school reforms: Effects of New American Schools in a high-poverty district*. Santa Monica, CA: RAND.
- Betts, E. A. (1946). *Foundations of reading instruction, with emphasis on differentiated guidance*. New York: American Book.
- Biggers, D. (2001). The argument against Accelerated Reader. *Journal of Adolescent & Adult Literacy*, 45, 72–75.
- Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80, 139–155.
- Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13, 4–16.
- Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive School Reform and achievement: A meta-analysis. *Review of Educational Research*, 73, 125–230.
- Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.
- Carter, B. (1996). Hold the applause!: Do Accelerated Reader and Electronic Bookshelf send the right message? *School Library Journal*, 42, 22–25.
- Cuddeback, M. J., & Ceprano, M. A. (2002). The use of Accelerated Reader with emergent readers. *Reading Improvement*, 39, 89–96.
- Cutler, V. (2002). *Reading proficiency more than doubles on Massachusetts Comprehensive Assessment System* (Renaissance Independent Research Rep. No. 55). Wisconsin Rapids, WI: Renaissance Learning.
- Facemire, N. E. (2000). The effects of the accelerated readers program on the reading comprehension of third graders. Unpublished master's thesis, Salem-Takeyo University, Salem, WV.
- Fine, A. (2001). *Alabama elementary school receives governor's trophy for most improvement after implementing Reading Renaissance* (Renaissance Independent Research Rep. No. 33). Wisconsin Rapids, WI: Renaissance Learning.
- Holmes, C. T., & Brown, C. L. (2002). *A controlled evaluation of a total school improvement process, School Renaissance*. Athens: University of Georgia.
- Howard, C. A. (1999). *An evaluation of the Accelerated Reader program in grades 3–5 on reading vocabulary, comprehension, and attitude in an urban southeastern school district in Virginia*. Unpublished doctoral dissertation, Old Dominion University, Norfolk, VA.
- Kambaran, V. N. (2001). *The role of reading instruction and the effect of a reading management system on at-risk students*. Unpublished doctoral dissertation, Saint Louis University, St. Louis, MO.
- Krashen, S. D. (2003). The (lack of) experimental evidence supporting the use of Accelerated Reader. *Journal of Children's Literature*, 29, 16–30.
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48, 1181–1209.
- Marzano, R. J. (2003). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Morris, S. (2001). *Georgia elementary school achieves growth in ITBS scores through Reading Renaissance implementation* (Renaissance Independent Research Rep. No. 7). Wisconsin Rapids, WI: Renaissance Learning.
- National Clearinghouse on Comprehensive School Reform. (2004). *The catalog of school reform models*. Retrieved August 7, 2005, from <http://www.nwrel.org/scpd/catalog>
- Nebelsick-Gullett, L. (2003). Review of STAR Reading version 2.2. In B. Flake, J. Impara, & R. Spies (Eds.), *The fifteenth mental measurements yearbook* (869–871). Lincoln, NE: Buros Institute.
- Paul, T. D. (2003). *Guided independent reading: An examination of the reading practice database and the scientific research supporting guided independent reading as implemented in Reading Renaissance*. Wisconsin Rapids, WI: Renaissance Learning.
- Pavonetti, L. M., Brimmer, K. M., & Ciplewski, J. F. (2000, November). *Accelerated Reader: What are the lasting effects on the reading habits of middle school students exposed in Accelerated Reader in elementary grades*. Paper presented at the annual meeting of the National Reading Conference, Scottsdale, AZ.
- Peak, J., & Dewalt, M. W. (1993, February). *Effects of the computerized Accelerated Reader program on reading achievement*. Paper presented at the annual meeting of the Eastern Educational Research Association, Clearwater Beach, FL.
- Sadusky, L. A., & Brem, S. K. (2002). *The integration of Renaissance programs into an urban Title I elementary school, and its effect on school-wide improvement*. Tempe: Arizona State University.
- Samuels, S. J., Lewis, M., Wu, Y., Reiningger, J., & Murphy, A. (2003). *Accelerated Reader vs. Non-Accelerated Reader: How students using the Accelerated Reader outperformed the control condition in a tightly controlled experimental study* (Tech. Rep.). Minneapolis: University of Minnesota.
- Samuels, S. J., & Wu, Y. (2003). *The effects of immediate feedback on reading achievement* (Tech. Rep.). Minneapolis: University of Minnesota.
- Scott, L. S. (1999). *The Accelerated Reader program, reading achievement, and attitudes of students with learning disabilities*. (ERIC Document Reproduction Service No. ED434431)
- Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- STAR Reading. (2001). *Understanding reliability and validity (Version 2.2)*. Wisconsin Rapids, WI: Advantage Learning Systems.
- Stevenson, J. M., & Camarata, J. W. (2000). Imposters in whole language clothing: Undressing the Accelerated Reader program. *Talking Points*, 11, 8–11.

- Terrance, P., VanderZee, D., Rue, T., & Swanson, S. (1996, October). *Impact of the Accelerated Reader technology-based literacy program on overall academic achievement and school attendance*. Paper presented at the National Reading Research Center Conference, Atlanta, GA.
- Topping, K. J., & Sanders, W. L. (2000). Teacher effectiveness and computer assessment of reading: Relating value added and learning information system data. *School Effectiveness and School Improvement, 11*, 305–337.
- Toro, A. (2001). *A comparison of reading achievement in second grade students using the Accelerated Reading program and independent reading*. Unpublished master of arts action research project, Johnson Bible College, Knoxville, TN.
- Vollands, S. R., Topping, K. J., & Evans, R. M. (1999). Computerized self-assessment of reading comprehension with the Accelerated Reader: Action research. *Reading & Writing Quarterly, 15*, 197–211.
- Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership, 41*, 19–27.
- Walberg, H. J. (1999). Productive teaching. In H. C. Waxman & H. J. Walberg (Eds.), *New directions for teaching practice and research* (pp. 75–104). Berkeley, CA: McCutchan.
- Wise, K. C., & Okey, J. R. (1983). A meta-analysis of the effects of various science teaching strategies on achievement. *Journal of Research in Science Teaching, 20*, 415–425.