Routledge
Taylor & Francis Group

Journal of Research on Educational Effectiveness

ISSN: 1934-5747 (Print) 1934-5739 (Online) Journal homepage: https://www.tandfonline.com/loi/uree20

Evaluating the Effectiveness of *Read Well Kindergarten*

Barbara Gunn , Keith Smolkowski & Patricia Vadasy

To cite this article: Barbara Gunn , Keith Smolkowski & Patricia Vadasy (2010) Evaluating the Effectiveness of *Read Well Kindergarten* , Journal of Research on Educational Effectiveness, 4:1, 53-86, DOI: 10.1080/19345747.2010.488716

To link to this article: https://doi.org/10.1080/19345747.2010.488716

Journal of Research on Educational Effectiveness, 4: 53–86, 2011

ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2010.488716

Routledge
Taylor & Francis Group

Evaluating the Effectiveness of *Read Well Kindergarten*

Barbara Gunn and Keith Smolkowski
Oregon Research Institute, Eugene, Oregon, USA

Patricia Vadasy
Washington Research Institute, Seattle, Washington, USA

Abstract: This article reports the outcomes of an experimental evaluation of *Read Well Kindergarten* (RWK), a program that focuses on the development of vocabulary, phonological awareness, alphabetic understanding, and decoding. Kindergarten teachers in 24 elementary schools in New Mexico and Oregon were randomly assigned, by school, to teach RWK or their own program. Treatment teachers received 2 days of training and taught daily lessons. Project staff assessed 1,520 students at pretest and 1,428 at posttest with measures of vocabulary, phonological awareness, alphabetic understanding, and decoding. Follow-up testing was conducted in fall and spring of first grade. Analyses of final outcomes revealed a statistically significant difference favoring intervention students on curriculum-based measures of sight words and decodable words. Although these results did not generalize to standardized measures, follow-up analyses indicated that the impact of RWK rested on the rate of opportunities for independent student practice for letter names, letter sounds, sight words, and oral reading fluency, collected at the end of kindergarten. The findings suggest the potential efficacy of RWK in conjunction with frequent opportunities for independent practice for developing beginning reading skills.

Keywords: Reading, kindergarten, curriculum, randomized trial

Reading is essential for success in school and fundamental for living in an information-based society. Although the reading skills of 4th- and 8th-grade students have increased, gaps between White and minority students have not been closed, and overall reading proficiency for 12th graders has declined (Lee, Grigg, & Donahue, 2007). Children who leave elementary school as poor readers frequently have difficulties learning to read in kindergarten and first grade and fail to establish a strong foundation of reading skills during their middle elementary school years (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Juel, 1988; Shaywitz & Shaywitz, 1999). This finding, now well substantiated, points to the compelling need to provide children with high-quality reading instruction in kindergarten and first grade.

The knowledge base on reading acquisition has grown considerably in the last decade, state and federal policies call for evidence-based instruction (Institute of Education Sciences, 2003), and independent research organizations (e.g., Coalition for Evidence-Based Policy, 2009; National Research Council, 2009) support rigorous research on educational interventions. In turn, publishers of commercially developed reading programs have responded by incorporating evidence-based practices in their curriculum design and content. Yet despite these advances, beginning reading programs still vary in quality, and primary

Address correspondence to Barbara Gunn, 1715 Franklin Boulevard, Oregon Research Institute, Eugene, OR 97403, USA. E-mail: barbarag@ori.org

teachers do not consistently receive adequate training and support to teach reading (Bursuck, Munk, Nelson, & Curran, 2002; Mather, Bos, & Babur, 2001; McCardle, Cooper, Houle, Karp, & Paul-Brown, 2001; McCutchen et al., 2002).

For the teacher or administrator looking for demonstrations of program effectiveness from randomized controlled trials, the evidence is sparse. In an examination of 806 articles from special education journals, Seethaler and Fuchs (2005) found that only 5.46% of the relevant articles in peer-reviewed journals tested a reading or math intervention using a group design and only 4.22% used random assignment. Our review of articles published in regular education journals indicated a similar pattern. Evaluations of reading interventions published since 1995 typically used quasi-experimental or descriptive designs, and those that used experimental designs to evaluate a well-defined curriculum had limited statistical power. Although this body of research contributes to the knowledge base by identifying promising approaches or interventions, the designs lack the internal validity to determine the relationship between intervention and student outcomes or to draw conclusions about overall program effectiveness.

PURPOSE OF THE STUDY

This study addresses the need for randomized field trials of reading programs to evaluate their effectiveness in teaching beginning reading skills. The primary aim of the project was to compare the effectiveness of *Read Well Kindergarten* (RWK) to the literacy instruction typically provided in kindergarten classrooms on the development of students' vocabulary, phonological awareness, alphabetic understanding, and decoding skills at the end of kindergarten and in the fall and spring of first grade. Among RWK classrooms we also examined whether student performance on outcome measures differed according to teachers' level of program implementation. Related aims of the study included tests of differential effects of 1 versus 2 years of implementation with RWK, school-level free and reduced lunch, and initial preliteracy scores. Finally, the study aimed to examine the interaction between rates of student independent practice and condition.

OVERVIEW OF THE INTERVENTION

RWK is a reading program designed for kindergarten children who are beginning readers and for primary grade students who need more instruction to develop beginning reading skills. The program is designed to develop (a) vocabulary, (b) phonological awareness, (c) alphabetic understanding, and (d) decoding (Biemiller, 1999; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Lonigan, Burgess, & Anthony, 2000; National Reading Panel, 2000; Torgesen, 2000). The program has a whole-class component that provides general exposure to early literacy concepts with read alouds, songs, and games and a small-group mastery-based component that is designed to provide students with explicit instruction in basic literacy skills.

RESEARCH ON KINDERGARTEN READING INSTRUCTION

Studies of kindergarten reading instruction have examined the feasibility and efficacy of interventions for children with typically developing reading skills and for those at risk

for reading difficulty. Building on the key role of instruction in phonological awareness and alphabetic understanding for developing beginning reading skills (Ball & Blachman, 1991; Bradley & Bryant, 1983, 1985), recent investigations of kindergarten reading instruction have sought to identify the optimal processes, components, and conditions of instruction for preventing reading difficulties and promoting the development of reading ability. Several researchers have investigated schoolwide implementation of commercially produced kindergarten reading programs that provide instruction in phonological awareness and alphabetic understanding. Al Otaiba et al. (2008) observed the implementation of evidence-based kindergarten reading programs in 17 *Reading First* schools. Their findings indicated that although all students received daily code-focused instruction and made gains in phonological awareness skills, they were not at grade level at the end of the year. Findings also indicated that children with lower initial levels of reading and vocabulary skills were more vulnerable to the quality of the instruction they received, suggesting the additional impact of instructional quality. Coyne, Kame'enui, Simmons, and Harn (2004) tracked the first-grade progress of children who received instruction with one of three kindergarten reading programs. They found that most students who made progress with the kindergarten programs did not need additional intervention in first grade to make comparable progress to their average-achieving peers. D'Angiulli, Siegel, and Maggi (2004) tracked English-speaking (L1) and English Language Learners (ELLs) who received instruction focused on reading and phonological processing beginning in kindergarten. By the end of Grade 5 the trajectory of literacy development for middle socioeconomic-status (SES) ELL students matched that of the L1 students. Among low and high SES students the ELLs improved more than the L1 students, suggesting that evidence-based kindergarten instruction may have attenuated the negative influence of SES on the development of decoding skills.

In addition to schoolwide studies of evidence-based kindergarten reading programs, researchers have also replicated findings on the specific contributions of instruction in phonological awareness and alphabetic understanding alone and in combination, with samples of at-risk kindergarten children (D. Fuchs et al., 2001; Hatcher et al., 2006; Hatcher, Hulme, & Snowling, 2004; Savage & Carless, 2005; Schneider, Roth, & Ennemoser, 2000; Vadasy, Sanders, & Peyton, 2006). Schneider et al. (2000) compared the effects of training in letter-sounds and phonological awareness alone and combined, delivered by kindergarten teachers to at-risk children. Children who received 5 months of combined training outperformed those in the other groups on measures of reading and spelling in Grades 1 and 2. Likewise, D. Fuchs et al. (2001) contrasted the effects of phonological awareness training with and without beginning decoding instruction delivered by kindergarten teachers to low-, average-, and high-achieving students and found improved outcomes for children receiving the combined instruction on measures of spelling and reading. Training studies with tutors or instructional assistants have yielded similar results for at-risk kindergarten students. Vadasy et al. (2006) assigned students to receive 18 weeks of instruction in phonemic skills and the alphabetic code or to a comparison condition. Significant differences emerged between groups on measures of spelling and reading at the end of the intervention and at 1-year follow-up. Hatcher et al. (2006) evaluated a 10- and 20-week intervention of letter-sound and phoneme identification activities delivered by teaching assistants to reading delayed children. Children who received the first 10 weeks of intervention initially made more progress on reading outcome measures. However, children who received only the second 10 weeks caught up to this group. About 25% of the children did not respond to the intervention, suggesting the need for a longer or more intensive intervention.

Collectively, the results of the kindergarten studies demonstrate that reading instruction that focuses on developing phonological awareness and alphabetic understanding can

impact reading outcomes for both normally developing and at-risk children. Findings also demonstrate that commercially produced programs can be used effectively by teachers and assistants, provided that instruction is explicit, intensive (Foorman & Torgesen, 2001), and focused on evidence-based processes.

RESEARCH ON *READ WELL*

To date, no experimental studies have been conducted on RWK. However, one quasi-experimental and several single-subject studies have examined the efficacy of *Read Well First Grade* (Sprick, Howard, & Fidanque, 1998), which uses the same content, instructional design, and mastery-based instruction as RWK. Findings have yielded mixed results on the program's effectiveness.

In a quasi-experimental study, Denton, Anthony, Parker, and Hasbrouck (2004) compared the efficacy of two English reading interventions for 93 Spanish-dominant ELLs in Grades 2 through 5. Students with low pretest scores were matched and randomly assigned to receive tutoring with *Read Well First Grade* or to a typical practice comparison group. Students with higher pretest scores were randomly assigned to receive *Read Naturally* (Ihnot, 1992), a program that uses repeated reading with contextualized vocabulary and comprehension instruction, or to typical school practice. Students were tutored three times weekly for 40 min over 10 weeks. Denton et al. (2004) compared the progress of tutored students ($n = 51$) and nontutored classmates ($n = 42$) using subtests of the Woodcock Reading Mastery Tests–Revised (WRMT–R). They reported a significant difference between students who received *Read Well* and the comparison group on the WRMT Word Identification subtest, with no significant effects on Word Attack or Passage Comprehension.

Jitendra et al. (2004) conducted two single-subject studies with 7 first-, second-, and third-grade students identified as having learning disabilities or attention deficit disorder, or as being ELLs. In Study 1, the students received 2 to 7 weeks of *Read Well First Grade* small-group instruction. Results indicated that three of the five students who received *Read Well* instruction improved in passage fluency, but their performance on the other measures of reading and comprehension was varied. The authors surmised that differences in student characteristics and the duration of *Read Well* instruction (2–7 weeks) may have accounted for differences in performance. In Study 2, Jitendra et al. implemented the *Read Well* intervention for up to 16 weeks with 5 second- and third-grade students identified by their teachers as poor readers. Two of the students had participated in Study 1. All five students made growth on measures of reading, spelling, and comprehension, with effect sizes greater than 2.0 on each measure.

In another single-subject design, Santoro, Jitendra, Starosta, and Sacks (2006) provided individual tutoring with *Read Well First Grade* to 4 second-grade ELLs who were poor readers. Students received an average of 30 min of instruction daily. Duration was 8, 14, 11, and 7 weeks and 16, 28, 22, and 14 hr, respectively, for the four students. Results indicated general improvement in word reading and oral reading fluency (ORF) for all students and improvement for two students on the WRMT Passage Comprehension subtest. Findings from these first-grade studies suggest the need for further investigation of *Read Well* using an experimental group design and a longer duration of implementation to determine the efficacy of the program for developing beginning decoding and comprehension skills for students with varying levels of early literacy skills.

SPECIFIC RESEARCH HYPOTHESES

The primary aim of this study was to compare the effectiveness of RWK to the literacy instruction typically provided in kindergarten classrooms on the development of students' vocabulary, phonological awareness, alphabetic understanding, and decoding skills at the end of kindergarten and in the fall and spring of first grade. We hypothesized that the specific instructional guidelines for teachers, controlled introduction of letters and sounds, cumulative review, and mastery-based approach in RWK would be more effective in developing beginning reading skills than the activities typically included in beginning reading programs (Coyne, Zipoli, & Ruby, 2006; Stein, Johnson, & Gutlohn, 1999; Stein, Stuen, Carnine, & Long, 2001).

We hypothesized that the benefits of RWK instruction would be maintained such that students in the RWK condition would outperform comparison students in the fall and possibly in the spring of first grade. Among RWK classrooms we also hypothesized that students of teachers with higher levels of implementation would outperform students of teachers with lower levels of implementation. Related aims tested differential or moderator effects of 1 versus 2 years of implementation with RWK, school-level free and reduced lunch, and risk of reading failure as indicated by pretest preliteracy scores.

The study also sought to examine the effects of independent student practice on the development of reading skills. We hypothesized that students would require sufficient practice to benefit from a curriculum and that students would benefit best from practice when taught with a curriculum based on research-validated content and approaches. That is, we hypothesized that both an evidence-based curriculum and an adequate amount of practice would be necessary for student learning but neither would be sufficient alone. Thus, we anticipated an interaction between treatment condition and the rate of independent student practice opportunities.

METHOD

Study Overview

The study included schools in both Oregon and New Mexico to test the effectiveness of RWK with a more diverse sample of students and teachers than might be found in either state alone. We recruited 26 elementary schools with kindergarten classrooms to participate in the study. The schools were randomly assigned to implement RWK or to a wait-list control condition. Two schools, 1 treatment and 1 control school, dropped out after randomization but before assessments, leaving 24 schools in the study. Participating kindergarten teachers at schools in the RWK condition received training from a certified *Read Well* trainer on how to teach the program with fidelity. Teachers in the control condition used the reading curriculum approved by their district and adopted by their school. We assessed all kindergarten children for whom we obtained parent consent in intervention and control classrooms on preliteracy and literacy measures before the intervention began and at the end of the kindergarten year. Students were assessed again in the fall and spring of first grade.

This study represents a group-randomized trial (Murray, 1998) with whole schools assigned to condition, intervention applied by teachers in intact classrooms, and measures collected on individual students nested within classrooms and schools. Because assignment to condition took place at the school level, statistical analysis modeled schools as the

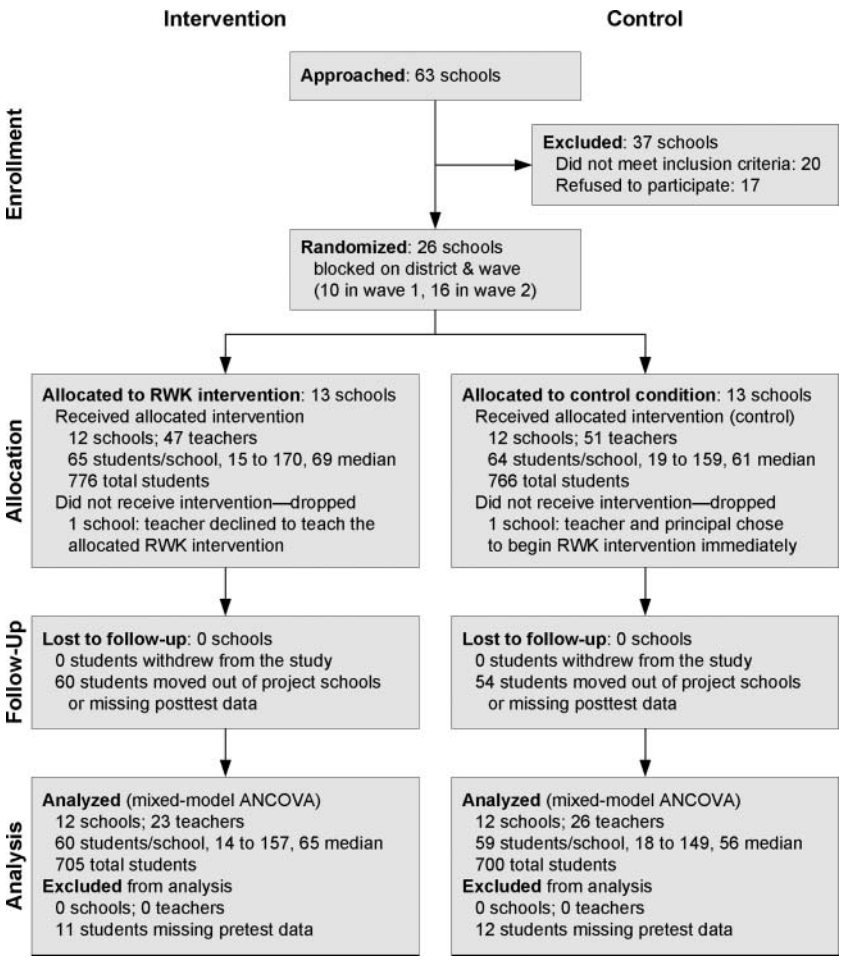

Figure 1. Research design and participant flow through *Read Well Kindergarten* randomized trial for schools and students. *Note*. Analysis sample sizes represent the mixed-model analysis of covariance (ANCOVA). The sample sizes for some measures differed by one to seven cases.

experimental unit. Explicit modeling of classrooms (subgroups) in tests of intervention effects was not necessary (Murray, Hannan, & Baker, 1996). The study was designed to have adequate power, based on analysis with a mixed-model analysis of covariance (ANCOVA; Murray, 1998), to detect small to medium effects of the intervention in kindergarten ($d =$ .41; Cohen, 1988). Figure 1 provides an overview of the research design and sample of students, teachers, and schools.

Participants

The study took place in 24 elementary schools, 12 in each intervention condition—19 in rural and suburban communities in Oregon and 5 in rural New Mexico. Overall, 54 teachers participated across the 3 school years, 26 in the RWK condition and 28 in the control condition, in the 24 schools. There were 37 intervention classes, 23 full day and 14 half

day. Of the 41 comparison classes, 23 were full-day classes and 18 were half-day classes. All kindergarten classrooms in New Mexico were full day. Four classrooms in Oregon were full day and the rest were either half day for a full week or full day for $2\frac{1}{2}$ days per week.

Within those classrooms, we assessed 1,519 kindergarten students in the fall and 1,427 in the spring. All students spoke English, although a few students spoke Spanish, Vietnamese, Mandarin, Arabic, or Tiwa (a Pueblo language) as their first language. Students in New Mexico were predominantly Hispanic, 60% to 80% by school, and the students receiving free or reduced lunch ranged from 65% to 94% by school. Oregon students were predominantly White, 70% to 85% from school to school, and the students receiving free or reduced lunch ranged from 20% to 60% by school. For the two Oregon schools that dropped out before intervention, 74% and 94% of the students were White, and 41% and 51% of students received free and reduced lunch, so they were within the range of the schools that remained in the project. Table 1 provides additional information about the sample of students, teachers, and schools, such as the percent of children on an individualized education plan.

Procedures

In the summer before the 1st year of the study, the principal investigator contacted superintendents in Oregon and New Mexico to describe the project and to ask permission to contact principals. Principals were sent an e-mail with an abstract of the study and then called. The principal investigator met with interested principals and kindergarten teachers to describe the project, explain the RWK program, and answer questions about their participation. To be eligible to participate, kindergarten teachers had to be using a reading program other than RWK at the time of recruitment and be willing to continue using either their current program or RWK after randomization. Teachers then made the final decision on whether to participate.

In the first year of the project (2005–2006), 10 schools, 4 in New Mexico and 6 in Oregon, were recruited and randomly assigned to the experimental (RWK small-group instruction) or control (regular reading program) condition. We refer to these schools as Wave 1, and they included 22 teachers. In the second year (2006–2007), we recruited and randomly assigned an additional 16 schools to condition. Because only 1 school joined the project in New Mexico, that school was combined with the Oregon schools for randomization. These Wave 2 schools included 26 more teachers. In the third year (2007–2008), another 6 teachers joined the study, for a total of 14 schools and 18 teachers.

Schools were blocked on district before random assignment to ensure balance within districts and states. After random assignment of Wave 2 schools but before assessment, 2 Oregon schools dropped out of the project, leaving 24 schools available for analysis. An intervention school withdrew because the teacher changed her mind and no longer wanted to teach RWK. A control school withdrew because the teacher and principal wanted to begin using RWK rather than wait 2 years when the study was completed.

Students were given baseline assessments after random assignment. Assessors were project research assistants or retired and substitute teachers who were unaware of students' intervention status. Assessors attended 2 days of standardized training prior to testing students. In addition, during the 1st day of testing in schools, the assessment coordinator first demonstrated how to test a student, then observed and gave assessors feedback on the fidelity and standardization of their testing procedures. For each subsequent assessment period, assessors attended a 4-hr refresher training on testing procedures and were observed

Table 1. Descriptive statistics for students, teachers, and schools by condition

Measure	*Read Well Kindergarten*	Control Schools
Students		
Age		
M	5.7	6.2
(*SD*)	(0.34)	(0.32)
Male	49.4%	50.9%
L1		
Spanish	3.3%	5.1%
Other	8.9%	6.5%
IEP		
Total	10.7%	11.5%
Speech & language	8.6%	8.6%
Autism	0.8%	0.7%
Specific learning disability	0.3%	0.9%
Developmental delay	0.6%	0.0%
Sample size	765	754
Teachers		
Years teaching		
M	12.6	11.6
(*SD*)	(10.12)	(10.10)
Years teaching K		
M	4.9	5.5
(*SD*)	(4.02)	(5.53)
Practice opportunities per minute		
M	7.6	4.1
(*SD*)	(2.90)	(2.64)
Sample size	26	28
Schools		
Free or reduced price lunch (%)		
M	52.2	54.6
(*SD*)	(14.63)	(18.02)
Sample size	12	12

Note. The sample sizes represent the maximum available. Age and sex were not available for 5 and 16 students, respectively. Language and individualized education plan (IEP) data were available for only 730 students in the RWK condition and 688 students in control classrooms. Teaching experience refers to the number of years in teachers' 1st year of participation. One RWK teacher did not provide data on teaching experience. L1 = first language.

again by the assessment coordinator during test administration to ensure standardization of their procedures. Students were tested in October before the intervention began and again in May.

To maintain integrity of data collection and mitigate potential conflict of interest, the principal investigator, one of the developers of RWK, was not associated with data collection activities, data management, or data analysis. The coinvestigator assumed primary responsibility for the oversight of the project assessments with the assistance of the assessment coordinator, who supervised the data collectors and facilitated transfer of data.

Reading Instruction

Treatment and comparison teachers conducted daily whole-class activities, such as the alphabet song and read alouds during their circle time. Treatment teachers provided basic literacy instruction to students grouped according to their instructional needs. Twenty-one of the 28 comparison teachers also provided basic literacy instruction to homogeneous groups of students. The remaining 7 teachers did not teach small groups but worked with children individually after the whole-class activities.

Intervention Schools. Teachers were asked to provide RWK instruction daily to students grouped according to their level of reading skill. The recommended instructional time was 20 to 25 min per group and the recommended group size was three to eight students. Summaries of classroom observations at six time points over 2 years indicated that intervention group sizes ranged from two to eight students with an average of 4.5 students per group. On average, teachers provided 23 min of instruction. The mean number of minutes per observation spent teaching specific content was 1.8 for phonological awareness, 9.2 for phonics, 0.20 for spelling, 1.1 for fluency, 7.5 for text reading, and 0.70 for comprehension. Instruction began after pretesting was completed in late October and continued until mid to late May. Eighty-nine percent of the teachers reported that they followed the RWK curriculum closely.

RWK Practice. Prior to beginning instruction, students were tested and placed at a level appropriate to their instructional needs. The daily lessons included (a) warm-up activities that reviewed previously taught sounds and words, (b) decoding practice with the student decoding magazines, and (c) story reading with decodable text. Small-group lessons included 6 Preludes (A–F) and 20 units. The Preludes prepared students for the more formal instruction in the units by focusing on basic concepts of print, phonological awareness, oral language, and comprehension skills. Each of the 20 units following the Preludes introduced a new letter-sound association and provided practice with decoding, fluency, and vocabulary. The instruction in the teacher guides was scripted. However, teachers were not required to use the script verbatim. Rather, they were asked to use the script to guide how they would teach the lessons.

RWK teachers spent 5 to 12 days on each Prelude or unit depending on how much time their students needed to learn the skills. Teachers gave students a curriculum-based test at the end of each Prelude or unit with specific criteria for progressing to the next unit. Students who did not pass the tests continued to receive instruction in their current unit until they could pass the test. Instruction focused on continued teaching of the same skills; the activities, however, were varied to keep students' interest and motivation. Individual students who did not pass by the end of the maximum days of instruction either repeated the unit with another small group or received extra individual practice with difficult skills.

RWK Training. In the 1st year of their participation, intervention teachers attended a standard 2-day publisher-developed training prior to teaching RWK. On Day 1, the trainer provided the research support for the program, an overview of the whole-class and small-group components, and a demonstration of whole-class instruction. On Day 2, teachers learned how to give the placement test, group students, and deliver small-group instruction. Day 2 training included trainer demonstrations and teacher practice, followed by feedback and discussions. Because many teachers had not used mastery-based instruction, the training emphasized the importance of teaching lessons with fidelity and using consistent

wording and clear instructions to minimize confusion for students. The principal investigator maintained regular phone and e-mail contact with the RWK teachers to answer their questions and to give them specific information on expected trajectories for their students. The principal investigator also made three coaching visits to each RWK classroom during the year to observe and give teachers feedback on their instruction and to demonstrate how to teach lessons with teachers observing. In the 2nd year of their participation, teachers attended an additional 1-day session on classroom management and a 1-day session focused on demonstration and practice of teaching routines, including more advanced skills such as correcting student errors and determining when to provide more practice and review. The principal investigator or the RWK trainer visited teachers three times each year, using the same coaching format as Year 1, and maintained regular phone and e-mail contact with them.

Comparison Schools. Across both years of the study, comparison teachers reported using commercially published reading programs for their main literacy instruction. The most commonly used were Harcourt Brace, McGraw-Hill, and Houghton Mifflin. The instructional focus of the activities in the programs was similar to the focus of the activities in the RWK classrooms. Summaries of classroom observations at six time points indicated that intervention small-group sizes ranged from two to eight students with an average of 5.7 students per group. On average, comparison teachers provided 24 min of instruction. The mean number of minutes per observation spent teaching specific content was 1.8 for phonological awareness, 6.8 for phonics, 2.4 for spelling, .20 for fluency, 2.5 for text reading, and 4.6 for comprehension. Forty percent of the teachers reported that they followed their curriculum closely.

Measures

Observations of Reading Instruction. Treatment and comparison teachers were observed by research staff at three time points during each year of their participation in the study. The 26 teachers in the RWK condition provided 113 observations and the 28 control teachers provided 122 observations. Observers used a multipage, scannable coding form to collect data on teachers' reading program, how students were grouped for instruction, the instructional focus and duration of activities, and four serially coded instructional behaviors observed during the reading lesson. A cover page also included more general information about the context, such as the number of students, data, and overall start and stop times.

Fidelity. For the three observations each year in RWK classrooms we also documented fidelity of implementation in two ways. First, observers documented on the observation form the RWK activity, such as Stretch and Shrink or Smooth and Bumpy Blending, referring to the plan for the lesson they were observing. For each activity observers then assigned a 3-point rating of fidelity. The fidelity for the activity was rated as 1 for low if the teacher did not follow any of the steps for the activity, 2 for medium if the teacher followed a minimum of two steps for the activity, and 3 for high if the teacher followed all the steps for the activity. We averaged those ratings across all activities and across the school year for each teacher. Second, observers provided a global rating on a scale of 1 to 3 for lesson implementation for each teacher, based on the degree to which teachers taught all the prescribed activities in each lesson.

Instructional Interactions. Observers also collected data on teacher–student instructional interactions. On the form observers coded each occurrence of (a) a teacher demonstration or model, (b) an independent student practice, (c) a student error, and (d) corrective feedback from the teacher. Because instructional interactions, particularly student independent practice, played a role in our original hypotheses we discuss this aspect of the observations in more detail.

The rationale for observing the specific behaviors of teachers and students during reading instruction was based on the research on instructional effectiveness and the role of teachers' instructional behavior in helping children learn new skills (e.g., Rosenshine, 1997; Vaughn, Gersten, & Chard, 2000). We were particularly interested in the effects of independent student practice on the development of reading skills, due to its theoretical support from the literature on instruction (e.g., Adams & Carnine, 2003; Foorman et al., 1998; D. Fuchs & Fuchs, 2005; Vellutino, Scanlon, Small, & Fanuele, 2006) and cognition (e.g., Shaywitz, Morris, & Shaywitz, 2008; Simos, Fletcher, Bergman, Breier, & Foorman, 2002; Stevens, Fanning, Coch, Sanders, & Neville, 2008).

We defined student independent practice as student practice of a skill, such as sounding out a word, or giving an answer to a question (e.g., "What sound does *M* make?") without any help from the teacher. To code independent practice, observers filled in one bubble on the scannable form each time during an activity in which either an individual or group of children practiced without any help from the teacher. Observers also coded the occurrence of teacher demonstrations, student errors, and teacher correction of errors using the same procedure to document where these interactions occurred in proximity to independent practice. For example, if students made an error during practice, observers coded whether the teacher corrected the error and then had students practice the skill again correctly. We computed rates per minute for each serial code to remove the influence of observation duration from the analyses, and for the analyses, we aggregated the rate of student practice at the school level. Teachers in RWK schools generally provided practice to students at a higher rate (see Table 1): 8 of 12 control schools and 6 of 12 RWK schools provided below-median rates of practice.

Training and Reliability. Trained project staff observed an entire literacy period, which was the time identified by teachers as their planned literacy instruction for the day. Observers were trained in four stages. First, they were given an overview of the system, an explanation of the codes, and procedures for using the observation codebook as a reference. In the second stage, the trainer and the observers practiced coding and debriefing observations as a group using video clips. In the third stage, the observers practiced coding with the trainer in kindergarten classrooms that were not in the study but that used RWK or other kindergarten curricula. In the fourth stage, reliability was established in project classrooms between the trainer and individual observers and maintained via periodic retesting. Observers who met and maintained 80% or higher rate of agreement with the trainer on the observations of instruction and RWK fidelity were allowed to observe project classrooms.

After collecting all observations, we calculated the interobserver reliability for the continuous rate of independent practice opportunities as an intraclass correlation coefficient (ICC; McGraw & Wong, 1996; Mitchell, 1979) and as the reliability of the observed mean (Raudenbush & Bryk, 2002). The ICC estimates the proportion of variance attributable to the target, the teacher in this case, opposed to observers, and it was .85. The reliability of the observed mean across multiple observations per classroom "measures the ratio of the *true score* or parameter variance, relative to the *observed score* or total variance of the

sample mean" (Raudenbush & Bryk, 2002, p. 46). This was calculated to be .81 in the present sample. Both statistics represent excellent reliability.

Student Measures. Project staff collected assessments at the beginning (T_1) and end (T_2) of kindergarten and the beginning (T_3) and end (T_4) of first grade. The measures included norm-referenced and criterion-referenced measures of alphabetic understanding, decoding, fluency, and receptive vocabulary. Table 2 provides descriptive statistics and shows the pattern of data collection for each measure and the student sample size at each assessment.

Rapid Automatized Naming. At pretest only, we collected rapid automatized naming of Objects, Colors, Numbers, Letter Naming, Letter and Number Naming, and Letters, Number and Color Naming subtests of the Rapid Automatized Naming/Rapid Automatized Stimulus tests (RAN/RAS; Wolf & Denckla, 2005). For each subtest, students are presented with a card that has the subtest item presented in random order and are asked to say the names of the target item, for example, colors, as quickly as they can. The raw score is the total number of seconds the student uses to name all of the colors. The test manual reported the test–retest reliability for each subtest in elementary grades as follows: Objects, .81; Colors, .86; Numbers, .89; Letters, .87; 2-set Letters and Numbers, .90; and 3-set Letters, Numbers, and Colors, .91.

Letter Names and Sounds. An untimed measure of letter names and sounds was administered at pretest and posttest. Students were shown a sheet with the letters of the alphabet in random order and asked to name the letters they knew. If students missed 5 letters consecutively, the examiner asked if they could name any other letters on the page. The score was the total number of letters named correctly, with a maximum score of 26 correct. The same procedure was followed for asking students to identify letter sounds. An untimed measure of letter identification was included in the test battery administered to kindergarten students in the Catts, Fey, Zhang, and Tomblin (2001) study of kindergarten predictors of early reading outcomes. Untimed measures of letter and word recognition have also been used in many other kindergarten studies of predictors of later reading outcomes (e.g., Bishop & League, 2006; Byrne, Fielding-Barnsley, Ashley, & Larsen, 1997; Muter, Hulme, Snowling, & Stevenson, 2004; Simpson & Everatt, 2005; Tunmer, Herriman, & Nesdale, 1988). Although assessors were carefully trained and monitored, reliability estimates were unavailable. As a proxy, we conducted test–retest correlations across the kindergarten school year as a lower bound indicator of reliability, but we used only control students ($n = 700$) due to the potential influence of the RWK intervention. Letter name and letter sound scores correlated .55 and .46 from T_1 to T_2.

Sight Words and Decodable Words. We used two untimed measures of word recognition to assess the near transfer effects of the intervention on word reading accuracy. The two untimed measures included sight words and decodable words taught in the phonics sequence of RWK. Sight words are words such as *was* or *said* that do not follow regular letter-sound relationships. Students were asked to read 10 high frequency sight words taken from the *American Heritage Word Frequency Book* (Carroll, 1971). Decodable words follow regular letter-sound relationships such as *had* or *seem* and can be sounded out using letter-sound correspondences. For this task, students were asked to read 12 words that contained sounds taught in RWK. Scores were reported as the number of words read correctly. Both tests were given at pretest and posttest. As with letter names and sounds, reliability estimates were unavailable, so we again conducted test–retest correlations across the kindergarten school

Table 2. Descriptive statistics for literacy measures by condition and assessment time

		Read Well Kindergarten				Control Schools			
Measure	Statistic	T_1	T_2	T_3	T_4	T_1	T_2	T_3	T_4
RAN	*M*	39.9				42.3			
	(SD)	(13.31)				(13.22)			
Letter names	*M*	16.8	24.5			17.4	24.1		
	(SD)	(9.29)	(3.73)			(8.82)	(4.17)		
Letter sounds	*M*	8.9	21.3			9.6	20.6		
	(SD)	(9.03)	(5.58)			(8.92)	(6.59)		
Sight words	*M*	1.4	6.8			1.8	5.4		
	(SD)	(2.12)	(3.08)			(2.42)	(3.13)		
Decodable words	*M*	1.0	7.9			1.2	5.1		
	(SD)	(2.48)	(3.56)			(2.89)	(4.10)		
CTOPP	*M*	2.4	5.0			2.4	5.0		
	(SD)	(2.81)	(3.85)			(2.57)	(3.85)		
PPVT	*M*	76.2	87.5		99.1	77.2	88.0		101.7
	(SD)	(17.68)	(18.01)		(17.79)	(17.77)	(17.44)		(17.06)
Word ID	*M*		11.7	16.1	38.8		12.0	17.0	40.5
	(SD)		(13.87)	(16.50)	(17.08)		(14.54)	(16.86)	(16.50)
Word attack	*M*		6.4	7.6	16.6		6.1	7.9	17.6
	(SD)		(6.71)	(7.54)	(9.48)		(6.88)	(7.97)	(9.50)
Oral reading fluency	*M*		15.8	14.5	46.6		15.1	16.3	51.7
	(SD)		(20.40)	(22.14)	(36.85)		(21.27)	(25.67)	(39.00)
Passage comprehension	*M*			7.48	17.91			7.31	19.28
	(SD)			(7.98)	(9.29)			(8.09)	(9.48)
Sample size	*N*	765	716	621	578	752	711	646	585

Note. The sample sizes represent the maximum students available across measures for each assessment period. The minimum sample size for any measure is at most 7 students fewer than the maximum. RAN = rapid automatized naming; CTOPP = Comprehensive Test of Phonological Processing; PPVT = Peabody Picture Vocabulary Test.

year with control students. Sight and decodable words correlated .46 and .43, respectively, from T_1 to T_2.

WRMT–R. The WRMT-R (Woodcock, 1998) is a standardized, individually administered battery of tests that measures multiple aspects of reading ability. The Word Identification subtest measures a student's ability to read sight words of increasing difficulty presented in isolation. The Word Attack subtest measures a student's ability to sound out and read a list of nonwords also presented in isolation. Because of the possible floor effects of these subtests in the fall of kindergarten, they were administered at posttest only. Internal consistency for the subtests ranges from .92 to .98.

ORF. ORF is a valid, reliable indicator of overall reading competence (L. S. Fuchs, Fuchs, Hosp, & Jenkins, 2001; Slocum, Street, & Gilberts, 1995; Stanovich, 2000). ORF rate was assessed at posttest in kindergarten with a beginning first-grade decodable reading passage entitled "Mac" (Maslen, 2003). We asked students to read a single passage in kindergarten. Students read the passage aloud for 1 min. Words omitted, substituted, and hesitations of more than 3 s are scored as errors. Words self-corrected within 3 s are scored as accurate. Fluency rate was calculated as the number of words correctly read in 1 min. Vadasy et al. (2006) reported internal consistency of .93 (26 words) at kindergarten posttest and .98 (121 words) at follow-up.

Fluency was also assessed in the fall and spring of first-grade students with three first-grade passages taken from the Dynamic Indicators of Basic Early Literacy Skills (Good & Kaminski, 2002) assessment battery. We used the same testing procedures as in kindergarten, except for the use of two additional passages. The analyses used the median score of the three passages. Baker et al. (2008) reported test–retest reliabilities of .94 or higher, and Good and Kaminski (2002) have shown alternate form reliabilities between .89 and .94.

Comprehensive Test of Phonological Processing (CTOPP). The CTOPP (Wagner, Torgesen, & Rashotte, 1999) is a normed, comprehensive measure of phonological processing. The Phoneme Elision subtest is a 20-item measure of the extent to which students can say a word and then say what remains after dropping out the designated sounds (i.e., either part of a compound word, part of an onset-rime, or an individual phoneme). The Elision subtest was given in the fall and spring of kindergarten to measure changes in student's phonological sensitivity. Raw scores were used for analysis to make comparisons between groups. Test–retest reliability as reported in the test manual is .88 for 5- to 7-year-olds.

We identified the CTOPP Phoneme Elision subtest as our measure of phonological processing specifically to include a measure of more advanced phonological processing, and to avoid ceiling effects. Our choice was also based on the Rosner and Simon (1971) study in which phoneme elision correlated with SAT reading, and provided kindergarten norms for this elision measure. Phoneme deletion was also included in the kindergarten screening procedures identified by Catts et al. (2001); Elbro, Borstrom, and Petersen (1998); Muter and Snowling (1998); and Muter et al. (2004) to identify predictors of second-grade reading outcomes. Muter (2000) found phoneme deletion the best predictor of reading skill at age 9 obtained at ages 5 to 6 years. Finally, CTOPP phoneme elision is one of the phonological processing tasks included in the widely used Texas Primary Reading Inventory kindergarten screening package.

Peabody Picture Vocabulary Test (PPVT). The PPVT (Dunn & Dunn, 1997) is an individually administered, norm-referenced, wide-range test of receptive language, or listening vocabulary. The test provides an estimate of verbal ability and the extent of English vocabulary acquisition. The test requires students to select one of four pictures that best illustrates the meaning of a stimulus word presented by the examiner. For this study, students were given Form IIIA at T_1, and Form IIIB at T_3. The PPVT has a mean of 100 and a standard deviation of 15. Test–retest reliability as reported in the test manual is .92 for 2- to 5-year-olds and .93 for 6- to 10-year-olds.

Analysis Methods

We assessed intervention effects on each of the primary outcomes with two different approaches, each with complementary strengths and weaknesses. For our primary analysis, we employed a mixed-model ANCOVA, but we also tested some measures with a nested time by condition (T × C) analysis. The second method provided a sensitivity analysis of the tenability of the underlying assumptions associated with each method (Greenland, 1996).

We first conducted a mixed-model ANCOVA with T_1 scores as covariates for T_2, T_3, and T_4 outcomes. The mixed-model ANCOVA nested students within schools, the unit of randomization. For measures collected at T_1 and at later assessments, we used the pretest score as covariates. For measures not collected at T_1, we added T_1 scores for sight words and decodable words as covariates. This analysis approach, then, contrasts residualized outcome scores between intervention and control conditions for students nested in schools. The model was expanded to test moderators and other interactions as well as tests of associations with mediators. In contrast, the nested T × C analysis was used to test differences between conditions on change in outcomes from T_1 to T_2 and parallels a repeated measures analysis with only two repeated assessments. This model also nested students within schools.

The mixed-model ANCOVA also generally offers greater power to detect differences between conditions (Janega et al., 2004a; Venter, Maxwell, & Bolig, 2002), although not always (Janega et al., 2004b). A disadvantage of this model, however, is that it includes only cases with data at both assessments. In contrast, the T × C analysis includes all data—whether or not a student's scores were present at both time points—to estimate differences between assessment points, and between conditions, which is generally more robust to effects of attrition. The T × C approach, however, assumes that "all groups and members respond to the intervention in the same way" (Murray, 1998, pp. 182–184), unlike the mixed-model ANCOVA, which "avoids altogether assumptions about the pattern of the within-member covariance matrices" (Murray, 1998, p. 190). The simple adjustment (ANCOVA) is also "less complex than the net difference in the repeated measures approach" (Murray, 1998, p. 190).

Because four measures were not collected at T_1 and because attrition was minimal (6%) between pretest and immediate posttest, T_1 to T_2, we used ANCOVA as our primary method of analysis and report the T × C analysis only when it led to different conclusions. Both analyses account for the ICC associated with multiple students nested within same schools. Although the design also nested students within classrooms, we assigned whole schools to condition; schools thus represented the unit of analysis. Murray et al. (1996) have shown that the analysis required for such a design need not include subgroups (e.g., classrooms)

to obtain the intended Type I error rate. To aid in planning similar group-randomized trials, we report ICC values from the mixed-model ANCOVA for each measure.

Model Estimation. We fit models to our data with SAS PROC MIXED version 9.1 (SAS Institute, 2004) using restricted maximum likelihood, generally recommended for multilevel models (Hox, 2002; Verbeke & Molenberghs, 2000). From each model, we estimated fixed effects and variance components. Maximum likelihood estimation for the T × C analysis also allows the use of all available data. Such an analysis gives unbiased results even in the face of substantial attrition, provided the missing data were missing at random (Laird, 1988; Schafer & Graham, 2002). In the present study, we did not believe that missing data represented a meaningful departure from the missing at random assumption, meaning that missing data did not likely depend on unobserved determinants of the outcomes of interest (Little & Rubin, 2002).

The estimated models assume independent and normally distributed observations. We addressed the first, more important assumption (van Belle, 2002) using multilevel statistical models. Regression and analysis of variance have also been found quite robust to violations of normality (Fitzmaurice, Laird, & Ware, 2004; Gibbons, Hedeker, Elkin, & Waternaux, 1993), and several studies have found that nonnormal data leads to acceptable results in a variety of multilevel modeling scenarios (Donner & Klar, 1996; Hannan & Murray, 1996; Maas & Hox, 2004a, 2004b; Murray et al., 2006). This feature of multilevel models eases concerns about the different scoring methods used for reading measures. The Woodcock measures, for example, provide several different scoring methods, including raw scores, *W* scores, normal curve equivalent scores, *T* scores, stanine scores, standardized scores, percentile ranks, and extended percentile scores. The literature does not clearly specify the most appropriate scoring method to use for research but concerns appear to revolve around the distributions. Because multilevel models provide robust results under many distribution assumptions, and because the different scoring methods have been shown to lead to very similar results in the multilevel analyses from randomized trials, the analyses of Woodcock and PPVT measures for the present study used raw scores.

Multiple tests, Type I and II errors, and effect sizes. In school-based randomized trials, the trade-off between Type I and II error rates represents a delicate balance. The cost of a false-positive conclusion, a Type I error, such as that the RWK curriculum improved student literacy skills when it in fact did not, is problematic. Yet with the high costs of school-randomized trials, false negatives or Type II errors can also obscure the value of potentially effective curricula. To balance these concerns, we recommend two ways to interpret statistical significance. First, with the usual criterion alpha of .05, we anticipate a 37% chance of one Type I error but only a 7.1% chance of two errors and a 0.8% chance of three or more Type I errors. We therefore interpret patterns of results for multiple items. For the interpretation of an individual test by itself, however, we recommend a criterion alpha of .006, which gives a 5.3% chance of one or more Type I errors.

To ease interpretation of effects we computed an effect size, Hedges's *g* (Hedges, 1981), for each fixed effect. Hedges's *g* represents an individual-level effect size comparable to Cohen's *d* (Rosenthal, Rosnow, & Rubin, 2000); Cohen's *d* uses a sample standard deviation and Hedges's *g* uses a population standard deviation (Rosenthal & Rosnow, 2008). The formula below transforms the coefficient for each modeled effect (γ) and the condition-specific sample sizes (n_R & n_C; where R denotes RWK and C control) and standard deviations (S_R & S_C) into Hedges' *g*, recommended by the What Works Clearinghouse

(2008) for multilevel models.

$$g = \frac{\gamma}{\sqrt{(n_R - 1)S_R^2 + (n_C - 1)S_C^2)/(n_R + n_C - 2)}}$$

Values for n_R, n_C, S_R, and S_C for main effects can be found in Table 2, with the values of γ reported in the Results section. Effect sizes for interactions use the same values for n_R, n_C, S_R, and S_C as condition main effects. The interpretation of g, however, changes for interactions with continuous variables, where it represents the effect size per unit change. For conditional effects that depend on moderators we report Hedges's g for the appropriate subsamples.

RESULTS

Table 2 provides means, standard deviations, and sample sizes for each literacy outcome by assessment time and condition. The analyses began, however, with an analysis of attrition, to test for differential student attrition effects. We then addressed each of our research hypotheses, including tests of main effects of condition, implementation fidelity, and moderation effects.

Attrition

Participant attrition, also called experimental mortality, poses a threat to both external and internal validity of a study (Barry, 2005; Graham & Donaldson, 1993; Shadish, Cook, & Campbell, 2002). We expected no attrition among schools, and none of the schools dropped from the study after assessments began. We also did not expect substantial attrition among students, nor did we expect student attrition to differ by treatment condition. Student attrition was calculated for T_2, T_3, and T_4, and it was defined as students missing data at a particular assessment and all subsequent assessments. Among the 776 total students in RWK schools and the 766 total students in comparison schools, the attrition rates were, respectively, 7.0% ($n = 54$) and 5.7% (44) at T_2, 19.5% ($n = 151$) and 15.0% ($n = 115$) at T_3, and 25.5% ($n = 198$) and 23.5% ($n = 180$) at T_4. At T_3, the beginning of first grade, we found that more RWK students had dropped from the study, $\chi^2(1) = 5.34$, $p = .0209$. Attrition rates did not differ at T_2 and T_4.

We next conducted an analysis to test whether student scores were differentially affected across conditions by attrition. The analysis examined the effects of condition, attrition status, and their interaction on pretest scores of letter names and sounds, sight and decodable words, CTOPP, and PPVT within a mixed-model analysis of variance. We tested for differential attrition on T_1 measures with T_2, T_3, and T_4 attrition variables and found no statistically significant interactions between attrition and condition.

Main Effects for RWK

This study aimed to test the hypothesis that students in RWK schools would outperform students in comparison schools by the end of kindergarten, with effects maintained into

first grade. As described in the Method section, the analyses included (a) a mixed-model analysis of covariance (ANCOVA) and (b) a T × C analysis. In nearly all cases the results from the T × C analysis confirmed those of the mixed-model ANCOVA, so we report only the results of the mixed-model ANCOVA. Table 3 presents complete model results for all kindergarten outcomes and includes ICC and Hedges's g values. We found no effects for first grade outcomes, with $p > .20$ in every case. For brevity, we omitted the results of those models.

Curriculum Based Measures. The analysis did not produce statistically significant main effects for letter names ($p = .1298$) or letter sounds ($p = .1547$). The main effect for condition statistically significantly favored the RWK condition for decodable words ($t = 4.67$, $p = .0001$, $g = 0.74$) as well as sight words ($t = 2.79$, $p = .0106$, $g = 0.46$). These curriculum-based measures assessed whether students learned what they were taught with the RWK curriculum. The next step, then, was to determine if the acquisition of these curriculum-based skills transferred to standardized outcome measures.

CTOPP. For the CTOPP, we found no main effects ($p = .6552$). The mean differences between conditions were small. From T_1 to T_2, students in the RWK conditions gained 2.60 correct responses and students in the control condition gained 2.57 correct responses, a difference of just .03. The mixed-model ANCOVA estimated an ICC for CTOPP of .017, which suggests that the school level accounts for less than 2% of CTOPP variance for kindergarteners. Results for the CTOPP were excluded from Table 3.

PPVT. Students in both conditions performed similarly on the PPVT by the end of kindergarten ($p = .6780$). We also collected the PPVT at the end of first grade (T_4), but found no main-effect differences ($p = .3651$). As shown in Table 2, the mean differences were negligible. From the mixed-model ANCOVA for main effects, we estimated ICC values of .022 at the end of kindergarten, indicating that schools account for only 2% of the variation in this measure of vocabulary. Due to the lack of significant effects, PPVT results were excluded from Table 3.

Word ID and Word Attack. Word ID and Word Attack were not administered at pretest (T_1), so the mixed-model ANCOVAs included sight words and decodable words as pretest covariates. We found no statistically significant condition effects for Word ID or Word Attack collected at the end of kindergarten (T_2, $p = .3787$, $p = .2640$, respectively) beginning of first grade (T_3, $p = .3304$, $p = .8661$) or end of first grade (T_4, $p = .7422$, $p = .6464$).

ORF. The ORF measure was administered at T_2, T_3, and T_4, so the ANCOVA models included sight words and decodable words as pretest covariates. We found no statistically significant main effects (T_2, $p = .3393$; T_3, $p = .8593$; T_4, $p = .5502$).

Passage Comprehension. We assessed students on passage comprehension in first grade, at T_3 and T_4. The ANCOVA models included sight words and decodable words collected at T_1 as covariates. The tests of main condition effects produced nonsignificant estimates at both T_3 ($p = .2296$) and T_4 ($p = .3483$). The ICC for passage comprehension from the main effects model at T_3 was .041. Passage comprehension results were not included in Table 3.

Table 3. Fixed effect (top) and variance component (bottom) estimates from a mixed-model analysis of covariance to test condition effects for read well kindergarten schools to comparison schools

Effect or Statistic	Letter Names	Letter Sounds	Sight Words	Decodable Words	Word ID	Word Attack	ORF
Fixed effects							
Intercept	20.24****	17.63****	4.40****	4.37****	5.51****	3.49****	6.61****
	(.28)	(.65)	(.36)	(.43)	(.86)	(.56)	(1.24)
Condition	.50	1.33	1.41*	2.85***	1.06	.90	1.66
	(.32)	(.90)	(.51)	(.61)	(1.18)	(.78)	(1.70)
Covariate (a)	.22****	.29****	.56****	.52****	2.54****	1.04****	3.19****
	(.01)	(.02)	(.03)	(.03)	(.22)	(.12)	(.33)
Covariate (b)					1.54****	.57****	2.71****
					(.19)	(.10)	(.28)
Variance components							
Residual	10.96****	25.57****	6.76****	10.98****	105.30****	29.03****	231.00****
	(.42)	(.97)	(.26)	(.42)	(4.01)	(1.11)	(8.81)
Teacher–Year	.34 ~	4.19**	1.35**	1.93**	5.76*	2.92*	11.55*
	(.18)	(1.47)	(.48)	(.69)	(2.59)	(1.14)	(5.19)
ICC	.030	.141	.166	.150	.052	.091	.048
Hedges's *g*							
Condition	.126	.218	.455	.743	.075	.132	.080

Note. Table entries show parameter estimates with standard errors in parentheses except for intraclass correlations (ICCs) and effect sizes (Hedges's *g*). Tests of the first four fixed effects (first four rows) used 22 *df*, tests of covariates used approximately 1375 *df*. For covariate (a), the analyses used the same measure collected at pretest (T_1) for Letter Names, Letter Sounds, Sight Words, and Decodable Words. Word ID, Word Attack, and Oral Reading Fluency (ORF) were not collected at pretest, so the analyses included pretest measures of (a) Sight Words and (b) Decodable Words as covariates.

$\sim p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $^{****}p < .0001$.

Implementation Fidelity in RWK Classrooms

We examined implementation fidelity with two measures: the mean activity-by-activity rating of implementation during our observations and the global rating of implementation by observers. The implementation ratings, averaged across all activities observed, ranged from 2.17 to 3.0 with a mean of 2.78 across all observations per teacher. The scores imply high levels of implementation fidelity with little variability. The analyses tested whether implementation fidelity could predict covariate adjusted outcomes, controlling for pretest performance, at the classroom level. Neither the mean rating from all individual activities nor the global rating predicted classroom-level gains in student literacy. Because the largest effect produced a correlation of only .26 ($p = .0821$), which represented only 6.8% overlapping variance between the global rating of implementation and the average T_2 CTOPP scores, controlling for T_1, we do not present the results for the remaining tests.

Implementation is also affected by the number of RWK units completed, as some teachers may have presented materials with fidelity but progressed too slowly or taught too infrequently for students to benefit. RWK teachers completed anywhere from 5 to 22.5 units during the school year out of 26 total units, with 22 teachers below the median (Unit 12) and 11 teachers below the 25th percentile (Unit 9). The number of units completed was correlated only .08 ($p = .5845$) with the global ratings of implementation and $-.02$ ($p = .9110$) with the mean rating of fidelity across activities. In contrast, the number of units completed was significantly associated with covariate-adjusted letter sounds ($r = .31$, $p = .0354$), sight words ($r = .71$, $p < .0001$), decodable words ($r = .57$, $p < .0001$), CTOPP ($r = .41$, $p = .0054$), PPVT ($r = .34$, $p = .0210$), Word ID ($r = .66$, $p < .0001$), Word Attack ($r = .66$, $p < .0001$), and ORF ($r = .72$, $p < .0001$) at the end of kindergarten (T_2). As elsewhere, the pretest covariates were either the same measure provided at pretest, if available, or letter names and sounds from pretest, when a pretest for the same measure was not available. In sum, these results demonstrate an association between outcomes and the frequency or quantity of instruction but not necessarily the fidelity of implementation.

Differential Effects by Implementation Year and School-Level Free and Reduced Lunch

This study involved two cohorts of students within each school. We therefore hypothesized that the students in the second cohort of RWK schools might perform better than those in the first cohort, due to teacher experience with the curriculum, whereas comparison students would remain more constant. To address this hypothesis, we tested the interaction between cohort and condition for all measures. No statistically significant effects were found.

It is also possible that schools with a higher average SES may outperform schools with lower SES and that this relationship may influence the success of a new curriculum. We tested the school-level proportion of students who receive free or reduced lunch as a moderator of condition effects but found no statistically significant interactions for the literacy measures.

Differential Effects by Risk of Reading Failure

Due to the design features of the RWK curriculum we hypothesized that students at risk of reading failure, those who perform more poorly on pretest predictors of reading, would

benefit most from RWK. To address this hypothesis, we tested whether the intervention differed between high and low initial scores on the RAN rate. RAN is a concurrent and longitudinal predictor of reading development (e.g., Denckla & Rudel, 1974; Kirby, Parrila, & Pfeiffer, 2003; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004; van den Bos, Zijlstra, & Spelberg, 2002) even among 5-year-olds (Simpson & Everatt, 2005). The RAN measure was chosen for these analyses to see if it would identify a subgroup of students less receptive to the intervention (see Johnston & Kirby, 2006) and to see if it would identify differences among subgroups in relation to rates of independent practice. Thus, the results presented next represent analyses that test Condition × RAN interactions. Table 4 presents the model estimates with RAN as a moderator.

Letter Names and Letter Sounds. RAN scores moderated the condition effects for letter names ($p = .0025$, $g = -0.01$, see Table 4 for details). The effect size, $g = -0.01$, indicates the change in the condition effect size for each unit change in RAN rate. For a 1-point decrease in RAN scores, students in the RWK condition performed 0.01 standard deviations better than controls. For letter names, students performed the same in each condition at the 79th percentile on RAN rates (score of 52.6), and the difference between conditions was statistically significant for students below the 45th percentile (38.0) on RAN rates at T_1 (Preacher, Curran, & Bauer, 2006). Among controls, students with RAN rates at the 25th and 75th percentiles (31.6 and 49.3) differed significantly on covariate adjusted letter name scores ($t = 4.29$, $p = .0003$). In contrast, the difference between students with RAN rates at the 25th and 75th percentiles was negligible within RWK schools ($t = -0.34$, $p = .7392$).

There was also a statistically significant interaction between RAN and condition for letter sounds ($p = .0062$, $g = -0.01$). Upon further inspection, the letter sound results favored the intervention condition among students with initial RAN rates below the 28th percentile. Students at the 75th percentile on initial RAN rates performed better than those at the 25th percentile in the RWK schools ($t = 3.03$, $p = .0061$) and even more so in control schools ($t = 7.17$, $p < .0001$) conditions.

Sight Words and Decodable Words. The initial RAN rate moderated condition effects for both covariate adjusted sight words ($p = .0380$, $g = -0.01$) and covariate adjusted decodable words ($p = .0017$, $g = -0.01$). Students in RWK schools statistically significantly outperformed students in comparison schools on sight words when they scored below the 83rd percentile on RAN (56.4) and on decodable words when they scored below the 96th percentile (71.2). The difference between students with initial RAN rates at the 25th and 75th percentiles was statistically significant for sight words within the RWK schools ($t = 6.04$, $p < .0001$) and control schools ($t = 9.01$, $p < .0001$) and for decodable words in RWK schools ($t = 5.61$, $p < .0001$) and control schools ($t = 10.55$, $p < .0001$).

CTOPP and PPVT. The analysis resulted in no statistically significant interactions with RAN rate for the CTOPP ($p = .0645$) or the PPVT ($p = .0843$) at the end of kindergarten or the PPVT at the end of first grade ($p = .5038$). Table 4 did omit the CTOPP and PPVT models.

Word ID and Word Attack. We found no interactions with initial RAN rate for either measure at the end of kindergarten (T_2; Word ID, $p = .6914$; Word Attack, $p = .3741$), at

Table 4. Fixed effect (top) and variance component (bottom) estimates from a mixed-model analysis of covariance to test moderation of condition by the risk for reading failure as indicated by the rate of RAN

Effect or Statistic	Letter Names	Letter Sounds	Sight Words	Decodable Words	Word ID	Word Attack	ORF
Fixed effects							
Intercept	20.49****	18.08****	4.60****	4.47****	6.29****	3.90****	7.82****
	(.29)	(.67)	(.37)	(.44)	(.87)	(.59)	(1.27)
Condition	.53	1.42	1.47**	2.95****	1.36	1.04	2.23
	(.32)	(.92)	(.51)	(.62)	(1.19)	(.82)	(1.74)
RAN Rate	.04***	.11****	.07****	.10****	.26****	.12****	.40****
	(.01)	(.02)	(.01)	(.01)	(.03)	(.02)	(.04)
RAN × Condition	−.05**	−.06**	−.02*	−.05**	−.02	.02	.04
	(.01)	(.02)	(.01)	(.01)	(.04)	(.02)	(.06)
Covariate (a)	.21****	.24****	.42****	.37****	2.06****	.81****	2.43****
	(.01)	(.02)	(.03)	(.03)	(.22)	(.11)	(.32)
Covariate (b)					1.40****	.49****	2.46****
					(.18)	(.09)	(.26)
Variance components							
Residual	10.39****	24.07****	6.22****	9.94****	96.57****	26.80****	207.01****
	(.40)	(.92)	(.24)	(.38)	(3.69)	(1.02)	(7.92)
Teacher–Year	.36 ∼	4.42**	1.40**	2.03**	5.99*	3.26**	12.76*
	(.19)	(1.53)	(.49)	(.72)	(2.68)	(1.26)	(5.58)
ICC	.034	.155	.184	.170	.058	.109	.058
Hedges's *g*							
RAN × Condition	−.011	−.010	−.007	−.012	−.001	.003	.002

Note. Table entries show parameter estimates with standard errors in parentheses except for intraclass correlations (ICCs) and effect sizes (Hedges' *g*). Tests of the first 4 fixed effects (first 4 rows) used 22 *df*, tests of covariates used approximately 1375 *df*. For covariate (a), the analyses used the same measure collected at pretest (T_1) for Letter Names, Letter Sounds, Sight Words, and Decodable Words. Word ID, Word Attack, and Oral Reading Fluency (ORF) were not collected at pretest, so the analyses included pretest measures of (a) Sight Words and (b) Decodable Words as covariates. RAN = rapid automatized naming.

$\sim p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $^{****}p < .0001$.

the beginning of first grade (T_3; Word ID, $p = .3304$; Word Attack, $p = .2816$) or at the end of first grade (T_4; Word ID, $p = .1781$; Word Attack, $p = .2179$).

ORF and Passage Comprehension. We found no statistically significant interactions with initial RAN rate for ORF at T_2 ($p = .5425$), T_3 ($p = .5898$), or T_4 ($p = .2177$). Tests of interactions between condition and pretest RAN rate were also not statistically significant for passage comprehension at T_3 ($p = .4730$) or T_4 ($p = .1111$).

Differential Effects by Rate of Independent Student Practice

Our final research hypothesis suggested that students require sufficient practice to benefit from the RWK curriculum. Stated differently, students benefit best from practice when coupled with the type of mastery-based, explicit instruction provided by the RWK curriculum. This hypothesis suggested that the interaction between condition and student independent practice opportunities may best predict student performance. Table 5 presents the results of these models; some models without significant results were excluded from the table.

Letter names and letter sounds. The analysis revealed a statistically significant interaction between condition and the rate of independent student practice opportunities per minute for both letter names ($p = .0348$, $g = 0.04$; see Table 5 for details) and letter sounds ($p = .0074$, $g = 0.06$). Although the interaction is statistically significant for letter names, this measure has no region of significance. That is, the confidence bounds for the condition effect include zero for the entire range of the rate of practice opportunities. For letter sounds, a statistically significant difference between conditions appeared at 6.9 opportunities per minute, the 98th percentile. See Figure 2 for graphs of measures with statistically significant interactions.

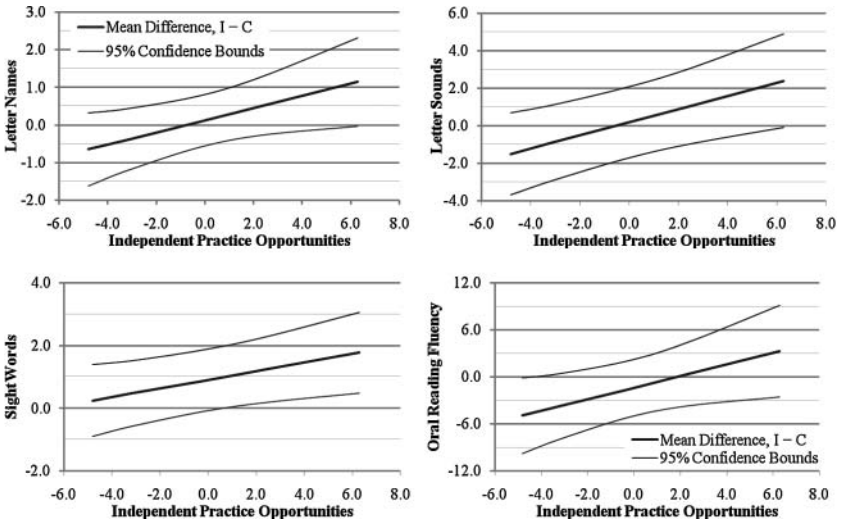

Figure 2. The graphs depict the interactions between treatment condition and independent student practice opportunities for letter names, letter sounds, decodable words, and oral reading fluency. *Note*. The vertical axis depicts the difference between condition on each dependent measure. The horizontal axis, for independent practice opportunities, was centered at the mean of 5.8.

Table 5. Fixed effect (top) and variance component (bottom) estimates from a mixed-model analysis of covariance to test interactions between treatment condition and rate per minute of independent student practice opportunities

Effect or Statistic	Letter Names	Letter Sounds	Sight Words	Decodable Words	Word ID	Word Attack	ORF
Fixed effects							
Intercept	20.24****	17.82****	4.54****	4.61****	6.40****	4.11****	7.45****
	(.30)	(.67)	(.34)	(.43)	(.85)	(.58)	(1.29)
Condition	.12	.18	.90 ∼	2.16**	−1.42	−.42	−1.40
	(.33)	(.91)	(.48)	(.60)	(1.15)	(.79)	(1.75)
Practice per minute	.02	.13	.07	.11 ∼	.43*	.29**	.42
	(.05)	(.09)	(.04)	(.06)	(.16)	(.09)	(.25)
Practice × Condition	.16*	.35**	.14*	.13	.40 ∼	.08	.74*
	(.07)	(.12)	(.06)	(.08)	(.23)	(.12)	(.34)
Covariate (a)	.22****	.30****	.57****	.53****	2.54****	1.04****	3.19****
	(.01)	(.02)	(.03)	(.03)	(.22)	(.12)	(.33)
Covariate (b)					1.57****	.58****	2.74****
					(.18)	(.10)	(.27)
Variance components							
Residual	10.89****	24.97****	6.66****	10.82****	103.25****	28.51****	227.12****
	(.41)	(.95)	(.25)	(.41)	(3.93)	(1.09)	(8.66)
Teacher–Year	.28 ∼	3.95**	1.08**	1.70**	4.21*	2.63*	10.09*
	(.17)	(1.42)	(.39)	(.61)	(1.98)	(1.03)	(4.56)
ICC	.025	.137	.140	.136	.039	.085	.043
Hedges's *g*							
Practice × Condition	.041	.058	.044	.035	.028	.012	.035

Note. Table entries show parameter estimates with standard errors in parentheses except for intraclass correlations (ICCs) and effect sizes (Hedges' *g*). Tests of the first 4 fixed effects (first 4 rows) used 22 *df*, tests of covariates used approximately 1375 *df*. For covariate (a), the analyses used the same measure collected at pretest (T_1) for Letter Names, Letter Sounds, Sight Words, and Decodable Words. Word ID, Word Attack, and Oral Reading Fluency (ORF) were not collected at pretest, so the analyses included pretest measures of (a) Sight Words and (b) Decodable Words as covariates.

$\sim p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $^{****}p < .0001$.

To understand the interaction, we compared students in schools with an average rate of practice at the 75th percentile (high-practice, 8.1 opportunities per minute) to those in schools at the 25th percentile (low-practice, 2.6 opportunities per minute) within each condition. Within the control condition, student adjusted letter name scores did not differ between high-practice and low-practice rates ($t = 0.51$, $p = .6146$), but within the RWK condition, letter name scores were significantly different between high and low rates of practice ($t = 3.58$, $p = .0071$). Similarly for letter sounds, the difference between high and low rates of practice was statistically significant among students in RWK schools ($t = 5.86$, $p < .0001$) but not controls ($t = 1.73$, $p = .0979$).

Sight words and decodable words. We found an interaction between condition and opportunities for student practice with sight words ($p = .0355$, $g = 0.04$) but not decodable words ($p = .1002$). Students provided with a rate of practice opportunities at the 60th percentile (6.6) or greater scored better on the sight words measure in RWK schools than control schools. Within the RWK schools, students scored better on sight words when provided greater opportunities for practice, at the 75th percentile, rather than fewer opportunities, at the 25th percentile ($t = 4.84$, $p < .0001$). The same is not true of control schools ($t = 1.53$, $p = .1394$).

CTOPP and PPVT. For the CTOPP, we found no interactions with practice opportunities ($p = .4932$). For the PPVT, we found no interactions between condition and the rate of opportunities for student practice at T_1 ($p = .2796$) or T_4 ($p = .9046$).

Word ID and Word Attack. The interaction between condition and the rate of practice was not significant for Word ID ($p = .0944$) or for Word Attack ($p = .5072$). We also found no statistically significant condition effects or interactions for Word ID or Word Attack collected at the beginning of first grade (T_3, $p = .7618$, $p = .6404$, respectively) or end of first grade (T_4, $p = .9778$, $p = .4008$).

ORF. The interaction between condition and the rate of practice was statistically significant for ORF ($t = 2.18$, $p = .0401$, $g = 0.04$). The region of significance for ORF included schools with a rate of practice below 1.3 opportunities per minute, below the 9th percentile, and in that narrow region, control schools outperformed RWK schools. In comparison schools, student ORF rates did not differ significantly when provided with high rates of practice, at the 75th percentile, versus low rates of practice, at the 25th percentile ($t = 1.67$, $p = .1091$). Within RWK schools, however, the difference in rate of practice produced a statistically significant difference in ORF scores ($t = 4.97$, $p < .0001$). The interaction between condition and the rate of practice was not statistically significant at T_3 ($p = .1192$) or T_4 ($p = .9785$).

Passage Comprehension. The interactions between condition and rate of student practice opportunities were not statistically significant at T_3 ($p = .2572$) or T_4 ($p = .8225$).

DISCUSSION

The primary aim of this study was to compare the effectiveness of RWK to the literacy instruction typically provided in kindergarten classrooms on the development of students' vocabulary, phonological awareness, alphabetic understanding, and decoding skills at the

end of kindergarten, and in the fall and spring of first grade. We also examined level of RWK implementation, differential effects of 1 versus 2 years of implementation with RWK, school-level free and reduced lunch, initial preliteracy scores, and independent practice opportunities. We summarize these results, present implications, describe the limitations of the study, and conclude with the significance and generalizability of the results in light of these factors.

Results Summary

We hypothesized that the specific instructional guidelines for teachers, controlled introduction of letters and sounds, cumulative review, and mastery-based approach in RWK would be more effective in developing beginning reading skills than the activities typically included in beginning reading programs and that these benefits would maintain through the fall and possibly the spring of first grade. Analyses of final outcomes revealed a statistically significant difference favoring intervention students on the curriculum-based measures of sight words and decodable words at the end of kindergarten, with no significant differences at either fall or spring follow-up assessments in first grade. The main effects for other measures, however, were not statistically significant. We were therefore unable to confirm benefit of RWK for students on standardized measures of sight word reading, decoding, phonemic awareness, and vocabulary as well as ORF.

Among RWK classrooms we anticipated that students of teachers with higher levels of implementation would outperform students of teachers with lower levels of implementation. Neither the mean rating from all individual activities nor the global rating predicted covariate adjusted classroom-level gains in student literacy. However, implementation quantity, as measured by the number of RWK units completed, was significantly associated with covariate-adjusted letter sounds, sight words, decodable words, CTOPP, PPVT, Word ID, Word Attack, and ORF at the end of kindergarten. We discuss these findings further under the Study Implications section.

Related aims tested differential or moderator effects of 1 versus 2 years of implementation with RWK, school-level free and reduced lunch, and risk of reading failure as indicated by pretest preliteracy scores. For these aims, we hypothesized that the students in the second cohort of RWK schools would perform better than those in the first cohort, due to teacher experience with the curriculum, whereas comparison students would remain more constant. No statistically significant effects were found. Similarly, we expected that schools with a higher than average SES might outperform schools with lower SES. Tests of the proportion of students who receive free or reduced lunch as a moderator of condition found no statistically significant interactions for any of the literacy measures.

With respect to risk, as measured by pretest RAN scores, we anticipated that students at greater risk for reading difficulty would benefit more from the instruction provided by RWK. We tested whether the intervention differences depended on initial scores on the RAN rate. Initial RAN rates moderated the difference between conditions for letter names, letter sounds, sight words, and decodable words. The difference between conditions was most apparent for students with low scores on initial RAN rates, with significant condition effects at the lower end of the RAN scale for all four measures. Within control schools, the difference between students with a high and a low initial RAN rate was statistically significant for all four measures, but not for letter names in RWK schools and to a lesser extent for letter sounds, sight words, and decodable words in RWK schools.

Finally, the study examined the interaction between rates of student independent practice and condition. We hypothesized that students would require sufficient practice to benefit from a curriculum, specifically in terms of learning beginning reading skills such as letter sounds and decoding, and that students would benefit best from practice when taught with a curriculum based on research-validated content and approaches. The analyses produced interactions between condition and rate of practice opportunities for four of the nine primary outcome measures collected at the end of kindergarten. We found neither statistically significant interactions nor main-effect differences for Word Attack, CTOPP, and PPVT collected in kindergarten, or Word Attack, Word ID, PPVT, and ORF collected during first grade (T_3 or T_4).

For letter names, letter sounds, sight words, and ORF, collected at the end of kindergarten, the impact of RWK rested on the rate of opportunities for independent student practice. For schools at or above the 60th percentile on the rate of independent practice opportunities (6.6 per minute), students exposed to RWK outperformed comparison students on sight words. Although high versus low rates of practice, the 25th versus the 75th percentile, predicted differences in each of the four outcomes within the RWK schools, we found no similar differences between students exposed to high versus low rates of practice within control schools. That is, within only the RWK schools did student gains depend on the rate of students' independent opportunities to practice.

Study Implications

Overall, our hypotheses received encouraging but not complete support. First, because of the statistically significant interactions between condition and independent opportunities for student practice, we must interpret the treatment condition effects conditional upon those interactions (Jaccard & Turrisi, 2003). The results, therefore, appear to support the position that exposure to the RWK curriculum alone is insufficient to produce student gains, as we found condition effects when ignoring the interactions for only curriculum-based sight words and decodable words. For letter names, letter sounds, sight words, and ORF, exposure to the curriculum produced greater gains when teachers also provided students with a high rate of independent practice opportunities. But practice, alone, also appeared insufficient, because we found no differences between high and low rates of practice opportunities within control schools. Students therefore appear to benefit from the RWK curriculum when delivered with sufficiently frequent opportunities for independent practice, which may suggest a need for more emphasis on training RWK teachers to provide repeated practice during instruction.

Kindergarten training studies support the effectiveness of systematic, explicit instruction for teaching phonological awareness and decoding skills (e.g., Ball & Blachman, 1991; Brady, Fowler, Stone, & Winbury, 1994; Wagner et al., 1997) and research on instructional intensity and skill acquisition supports the need to practice new skills until they are fluent and automatic. In light of our findings the study suggests that the combination of instruction with RWK and frequent opportunities for independent practice may be efficacious for kindergarten students; however, further experimental research is needed to verify the short- and long-term effects of instruction with the complete scope and sequence of the program.

RWK appeared to mitigate differences between students who performed more poorly and those who performed better on rapid automatized naming. The students with lower RAN rates made greater improvements in RWK schools than in controls, and we found larger differences in gains between students with low and high RAN rates, defined by

the 25th versus the 75th percentile, in the control condition than in the RWK condition. This may be a result of the combination of RWK and the additional practice. The present study, however, lacked the sample to test three-way interactions, the subsequent two-way conditional interactions, and conditional "main" effects (Jaccard & Turrisi, 2003). The analysis of such complicated models with a relatively small sample of classrooms and schools may well have led to sample-specific results (Burnham & Anderson, 2002; Myung, 2000; Zucchini, 2000), so we leave these questions to future research.

We measured aspects of RWK implementation with four methods. First, observers coded collected activity-by-activity ratings of implementation fidelity, and we averaged the multiple ratings per observation and multiple observations throughout the year. Second, observers provided a global impression of RWK implementation fidelity. Third, we documented the number of RWK units teachers completed with each group of students. Finally, we assessed two cohorts of students, assuming that teachers may improve their instruction after gaining familiarity with the curriculum, which we would characterize as a loose measure of fidelity. We found no relationship between fidelity ratings or cohort membership and any of the literacy outcomes. Fidelity ratings, however, were also generally high. We conclude, then, that either (a) fidelity was sufficient in (nearly) all classrooms or (b) the fidelity measures lacked sufficient variability to discriminate students based on their literacy scores.

Of interest to us was the finding on the significant relationship between number of RWK units completed and student outcomes. As noted, the number of units completed was not strongly correlated with the global or within-lesson ratings of fidelity. Although teachers agreed to teach RWK daily we observed that competing demands on their instructional time and, in some cases, a lack of understanding of the importance for daily instruction meant that some teachers did not teach a complete lesson every day. Thus, some may have presented materials with fidelity but progressed too slowly or taught too infrequently for students to benefit.

Limitations

This was a large-scale efficacy trial in which treatment interacted with independent practice. Only two of the nine outcome measures showed a main effect without moderators. We acknowledge that differences by condition on these two curriculum-based measures of sight words and decodable words do not necessarily support the effectiveness of the curriculum on the development of generalized reading skills. As noted, the purpose of the Curriculum Based Measures was to first assess whether students learned the skills they were taught with the RWK curriculum, which it appears they did. The next step then was to assess whether acquisition of these curriculum-based skills transferred to standardized outcome measures. Because there were no significant differences between condition on the standardized measures it appears that RWK was no more effective in improving student outcomes than the comparison approaches when only considering main effects. On average, RWK students learned 12 to 13 letter-sound correspondences, based on an average of 12.5 out of 26 RWK units completed per teacher. Had they learned the complete sequence of letter sounds taught in the curriculum their performance on standardized measures may have improved. However, the same may be said for instruction in comparison classrooms. Future experimental studies, which document the exact scope and sequence of instructional content across conditions, could better answer this question. In light of these limitations, and given that one of the authors is a program developer, we recommend and exercise caution in reporting and interpreting significant findings.

This study suffered from a few key differences between design and execution. We made, in hindsight, a few overly liberal assumptions, which may have reduced statistical power. First, we planned the study for 30 schools but, due to recruitment difficulties, obtained 24. Second, we anticipated three classrooms per school, but our schools contained, on average, only two participating teachers. Finally, the best evidence available during the research design phase indicated that an ICC of .04 was sufficient. This estimate came from our earlier research, but it was considerably lower than many of the ICCs estimates reported herein and estimates from recent sources (e.g., Hedges & Hedberg, 2007) that were not yet available when we designed this project. Combined, these differences between our study design and its realization substantially increased the smallest effect sizes that the analyses could detect. The results of this study allow for some conclusions, but we stress caution about liberal assumptions during study design. These can potentially lead to substantially reduced power after data have been collected and could severely limit the ability to adequately test important hypotheses.

Summary and Future Directions

We believe this study demonstrates the feasibility of conducting a randomized trial of an educational intervention in real educational settings. At a time when experimental evaluations of commercially produced curricula are still rare it is important to evaluate programs in order to provide educators with objective information on program effectiveness. At the same time it is important to study the conditions under which programs succeed to move beyond general recommendations and give teachers and administrators specific information on the critical roles of both program content and instructional practices for ensuring optimal outcomes. In sum, the present study suggests that the combination of an evidence-based reading program and sufficient teacher–student interactions—specifically independent practice opportunities—can lead to significant gains in literacy for kindergarten students.

ACKNOWLEDGEMENTS

This research was supported by the U.S. Department of Education, Institute of Education Sciences, grant R305F050080. Opinions expressed in this paper represent those of the authors and not the institutions with which they are affiliated or the U.S. Department of Education.

REFERENCES

Adams, G., & Carnine, D. (2003). Direct instruction. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 403–416). New York: Guilford.

Al Otaiba, S., Connor, C. M., Kosanovich, M., Schatschneider, C., Dyrlund, A. K., & Lane, H. (2008). Reading First kindergarten classroom instruction and students' growth in phonological awareness and letter naming-decoding fluency. *Journal of School Psychology*, *48*, 281–314.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., et al. (2008). Reading fluency as a predictor of reading proficiency in low performing high poverty schools. *School Psychology Review*, *37*(1), 18–37.

Ball, E. W., & Blachman, B. A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly*, *26*, 49–66.

Barry, A. E. (2005). How attrition impacts the internal and external validity of longitudinal research. *Journal of School Health*, *75*, 267–270.

Biemiller, A. (1999). *Language and reading success* (Vol. 5). Cambridge, MA: Brookline.

Bishop, A. G., & League, M. B. (2006). Identifying a multivariate screening model to predictreading difficulties at the onset of kindergarten: A longitudinal analysis. *Learning Disability Quarterly*, *29*, 235–252.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read—a causal connection. *Nature*, *301*, 419–421.

Bradley, L., & Bryant, P. E. (1985). *Rhyme and reason in reading and spelling*. Ann Arbor: University of Michigan Press.

Brady, S., Fowler, A., Stone, B., & Winbury, N. (1994). Training phonological awareness: A study with inner-city kindergarten children. *Annals of Dyslexia*, *44*, 26–59.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach* (2nd ed.). New York: Springer-Verlag.

Bursuck, W. D., Munk, D. D., Nelson, C., & Curran, M. (2002). Research on the prevention of reading problems: Are kindergarten and first grade teachers listening? *Preventing School Failure*, *47*(1), 4–9.

Byrne, B., Fielding-Barnsley, R., Ashley, L., & Larsen, K. (1997). Assessing the child's and the environment's contribution to reading acquisition: What we know and what we don't know. In B. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention Hillsdale* (pp. 265–285). Mahwah, NJ: Erlbaum and Associates.

Carroll, J. B. (1971). *The American heritage word frequency book*. Boston: Houghton Mifflin.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools*, *32*, 38–50.

Coalition for Evidence-Based Policy. (2009). *Evidence-based reform: A key to major gains in education, poverty reduction, crime prevention, health care, other areas*. Retrieved August 1, 2009, from http://coalition4evidence.org/wordpress/?page_id=6

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

Coyne, M. D., Kame'enui, E. J., Simmons, D. C., & Harn, B. A. (2004). Beginning reading intervention as inoculation or insulin; First-grade reading performance of strong responders to kindergarten intervention. *Journal of Learning Disabilities*, *37*, 90–104.

Coyne, M. D., Zipoli, R., & Ruby, M. (2006). Beginning reading instruction for students atrisk for reading disabilities: What, how, when. *Intervention in School & Clinic*, *41*, 161–168.

D'Anguilli, A., Siegel, L. S., & Maggi, S. (2004). Literacy instruction, SES, and word-reading achievement in English-language learners and children with English as a first language: A longitudinal study. *Learning Disabilities Research & Practice*, *19*, 202–213.

Denckla, M. B., & Rudel, R. (1974). Rapid "automatized" naming of pictured objects, colors, letters, and numbers by normal children. *Cortex*, *10*, 186–202.

Denton, C. A., Anthony, J. L., Parker, R., & Hasbrouck, J. (2004). Effects of two tutoring programs on the English reading development of Spanish-English bilingual students. *The Elementary School Journal*, *104*, 289–305.

Donner, A., & Klar, N. (1996). Statistical considerations in the design and analysis of community intervention trials. *Journal of Clinical Epidemiology*, *49*, 435–439.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Elbro, C., Borstrom, I., & Petersen, D. K. (1998). Predicting dyslexia from kindergarten: The importance of distinctiveness of phonological representations of lexical items. *Reading Research Quarterly*, *33*, 36–60.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.

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*(1), 37–55.

Foorman, B. R., & Torgesen, J. (2001). Critical elements of classroom and small group instruction promote reading success in all children. *Learning Disabilities Research & Practice*, *16*, 203–212.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, *88*(1), 3–17.

Fuchs, D., & Fuchs, L. S. (2005). Peer-assisted learning strategies: Promoting word recognition, fluency, and reading comprehension in young children. *Journal of Special Education*, *39*, 34–44.

Fuchs, D., Fuchs, L. S., Thompson, A., Al Otaiba, S., Yen, L., Yang, N. J., et al. (2001). Is reading important in reading-readiness programs? A randomized field trial with teachers as program implementers. *Journal of Educational Psychology*, *93*, 251–267.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*, 239–256.

Gibbons, R. D., Hedeker, D. R., Elkin, I., & Waternaux, C. (1993). Some conceptual and statistical issues in analysis of longitudinal psychiatric data: Application to the NIMH treatment of Depression Collaborative Research Program dataset. *Archives of General Psychiatry*, *50*, 739–750.

Good, R., & Kaminski, R. (2002). *DIBELS Oral Reading Fluency passages for first through third grades* (Tech. Rep. No. 10). Eugene: University of Oregon.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*, *78*, 119–128.

Greenland, S. (1996). Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology*, *25*, 1107–1116.

Hannan, P. J., & Murray, D. M. (1996). Gauss or Bernoulli? A Monte Carlo comparison of the performance of the linear mixed-model and the logistic mixed-model analysis in simulated community trials with a dichotomous outcome variable at the individual level. *Evaluation Review*, *20*, 338–352.

Hatcher, P. J., Hulme, C., Miles, J. N. V., Carroll, J. M., Hatcher, J., Gibbs, S., et al. (2006). Efficacy of small group reading intervention for beginning readers with reading-delay: A randomized controlled trial. *Journal of Child Psychology and Psychiatry*, *47*, 820–827.

Hatcher, P. J., Hulme, C., & Snowling, M. J. (2004). Explicit phoneme training combined with phonic reading instruction helps young children at risk of reading failure. *Journal of Child Psychology and Psychiatry*, *45*, 338–358.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87.

Hox, J. (2002). *Multilevel analysis techniques and applications*. Mahwah, NJ: Erlbaum.

Ihnot, C. (1992). *Read naturally*. St Paul, MN: Read Naturally.

Institute of Education Sciences. (2003, December). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: National Center for Education Evaluation and Regional Assistance, U.S. Department of Education.

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression* (2nd ed.). Thousand Oaks, CA: Sage.

Janega, J. B., Murray, D. M., Varnell, S. P., Blitstein, J. L., Birnbaum, A. S., & Lytle, L. A. (2004a). Assessing intervention effects in a school-based nutrition intervention trial: Which analytic model is most powerful? *Health Education and Behavior*, *31*, 756–774.

Janega, J. B., Murray, D. M., Varnell, S. P., Blitstein, J. L., Birnbaum, A. S., & Lytle, L. A. (2004b). Assessing the most powerful analysis method for school-based intervention studies with alcohol, tobacco, and other drug outcomes. *Addictive Behaviors*, *29*, 595–606.

Jitendra, A. K., Edwards, L. L., Starosta, K., Sacks, G., Jacobson, L. A., & Choutka, C. M. (2004). Early reading instruction for children with reading difficulties: Meeting the needs of diverse learners. *Journal of Learning Disabilities*, *37*, 421–439.

Johnston, T. C., & Kirby, J. R. (2006). The contribution of naming speed to the simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, *19*, 339–361.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, *80*, 437–447.

Kirby, J. R., Parrila, R., & Pfeiffer, S. (2003). Naming speed and phonological awareness as predictors of reading development. *Journal of Educational Psychology*, *95*, 453–464.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, *7*, 305–315.

Lee, J., Grigg, W., & Donahue, P. (2007). *The Nation's Report Card: Reading 2007* (NCES 2007–496). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

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

Lonigan, C. J., Burgess, S. R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent variable longitudinal study. *Developmental Psychology*, *36*, 596–613.

Maas, C. J. M., & Hox, J. J. (2004a). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*, 427–440.

Maas, C. J. M., & Hox, J. J. (2004b). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137.

Maslen, B. L. (2003). *Bob books*. New York: Scholastic.

Mather, N., Bos, C., & Babur, N. (2001). Perceptions and knowledge of preservice and inservice teachers about early literacy instruction. *Journal of Learning Disabilities*, *34*, 472–482.

McCardle, P., Cooper, J. A., Houle, G. R., Karp, N., & Paul-Brown, D. (2001). Emergent and early literacy: current status and research directions—introduction. *Learning Disabilities Research & Practice*, *16*, 183–185.

McCutchen, D., Abbott, R. D., Green, L. B., Beretvas, S. N., Cox, S., Potter, N. S., et al. (2002). Beginning literacy: Links among teacher knowledge, teacher practice, and student learning. *Journal of Learning Disabilities*, *35*(1), 69–86.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86*, 376–390.

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York. Oxford University Press.

Murray, D. M., Hannan, P. J., & Baker, W. L. (1996). A Monte Carlo study of alternative responses to intraclass correlation in community trials. *Evaluation Review*, *20*, 313–337.

Murray, D. M., Hannan, P. J., Pals, S. P., McCowen, R. G., Baker, W. L., & Blitstein, J. L. (2006). A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Statistics in Medicine*, *25*, 378–388.

Muter, V. (2000). Screening for early reading failure. In N. A. Badian (Ed.), *Prediction and prevention of reading failure* (pp. 1–30). Baltimore, MD: York Press.

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology*, *40*, 665–681.

Muter, V., & Snowling, M. J. (1998). Concurrent and longitudinal predictors of reading: The role of metalinguistic and short-term memory skills. *Reading Research Quarterly*, *33*, 320–337.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Pub. No. 00-4769). Washington, DC: U.S. Government Printing Office, Department of Health & Human Services. National Institute of Health. Available from http://www.nationalreadingpanel.org

National Research Council. (2009). *Division of behavioral and social sciences and education: Education*. Retrieved August 1, 2009, from http://sites.nationalacademies.org/NRC/

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rosenshine, B. (1997). Advances in research on instruction. In J. W. Lloyd, E. J. Kame'enui, & D. Chard (Eds.), *Issues in educating students with disabilities* (pp. 197–221). Mahwah, NJ: Erlbaum.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis* (3rd ed.). McGraw–Hill.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York: Cambridge University Press.

Rosner, J., & Simon, D. P. (1971). The auditory analysis test: An initial report. *Journal of Learning Disabilities*, *4*, 384–392.

Santoro, L. E., Jitendra, A. K., Starosta, K., & Sacks, G. (2006) Reading well with Read Well: Enhancing the reading performance of English language learners. *Remedial and Special Education*, *27*, 105–115.

SAS Institute. (2004). *SAS/STAT ® 9.1 user's guide*. Cary, NC: Author.

Savage, R., & Carless, S. (2005). Learning support assistants can deliver effective reading interventions for 'at-risk' children. *Educational Research*, *47*, 45–61.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R.,(2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, *96*, 265–282.

Schneider, W., Roth, E., & Ennemoser, M. (2000). Training phonological skills and letter knowledge in children at risk for dyslexia: A comparison of three kindergarten intervention programs. *Journal of Educational Psychology*, *92*, 284–295.

Seethaler, P. M., & Fuchs, L. S. (2005). A drop in the bucket: Randomized controlled trials testing reading and math interventions. *Learning Disabilities Research & Practice*, *20*, 98–102.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York: Houghton Mifflin.

Shaywitz, S., Morris, R., & Shaywitz, B. (2008). The education of dyslexic children from childhood to young adulthood. *The Annual Review of Psychology*, *59*, 451–475.

Shaywitz, S., & Shaywitz, B. (1999). Cognitive and neurobiologic influences in reading and in dyslexia. *Developmental Neuropsychology*, *16*, 383–384.

Simos, P. G., Fletcher, J. M., Bergman, E., Breier, J. I., & Foorman, B. R. (2002). Dyslexia-specific brain activation profile becomes normal following successful remedial training. *Neurology*, *58*, 1203–1213.

Simpson, J., & Everatt, J. (2005). Reception class predictors of literacy skills. *British Journal of Educational Psychology*, *75*, 171–188.

Slocum, T. A., Street, E. M., & Gilberts, G. (1995). A review of research and theory on the relation between oral reading rate and reading comprehension.*Journal of Behavioral Education*, *5*, 377–398.

Sprick, M., Howard, L. M., & Fidanque, A. (1998). *Read Well First Grade*. Longmont, CO: Sopris West Educational Services.

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: Guilford.

Stein, M., Johnson, B., & Gutlohn, L. (1999). Analyzing beginning reading programs. The relationship between decoding instruction and text. *Remedial and Special Education*, *20*, 275–288.

Stein, M., Stuen, C., Carnine, D., & Long, R. M. (2001). Textbook evaluation and adoption practices. *Reading and Writing Quarterly: Overcoming Learning Difficulties*, *17*(1), 5–23.

Stevens, C., Fanning, J., Coch, D., Sanders, L., & Neville, H. (2008). Neural mechanisms of selective auditory attention are enhanced by computerized training: Electrophysiological evidence from language-impaired and typically developing children. *Brain Research*, *1205*, 55–69.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice*, *15*(1), 55–64.

Tunmer, W. E., Herriman, M. L., & Nesdale, A. R. (1988). Metalinguistic abilities and beginning reading. *Reading Research Quarterly*, *23*, 134–158.

Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006). Code-oriented instruction for kindergarten students at risk for reading difficulties: A randomized field trial with paraeducator implementers. *Journal of Educational Psychology*, *98*, 508–528.

van Belle, G. (2002). *Statistical rules of thumb*. New York: Wiley.

van den Bos, K. P., Zijlstra, B. J. H., & Spelberg, H. C. (2002). Life-span data on continuous-naming speeds of numbers, letters, colors, and pictured objects, and word-reading speed. *Scientific Studies of Reading*, *6*, 25–49.

Vaughn, S., Gersten, R., & Chard, D. J. (2000). The underlying message in LD intervention research: Findings from research syntheses. *Exceptional Children*, *67*, 99–114.

Vellutino, F. R., Scanlon, D. M., Small, S., & Fanuele, D. P. (2006). Response to intervention as a vehicle for distinguishing between children with and without reading disabilities: Evidence for the role of kindergarten and first-grade interventions. *Journal of Learning Disabilities*, *39*, 157–169.

Venter, A., Maxwell, S. E., & Bolig, E. (2002). Power in randomized group comparisons: The value of adding a single intermediate time point to a traditional pretest-posttest design. *Psychological Methods*, *7*, 194–209.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

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

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., et al. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology*, *33*, 468–479.

What Works Clearinghouse. (2008). *Procedures and standards handbook (*Version 2.0*)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved from the Institute of Education Sciences, National Center for Education Evaluation, WWC: http://ies.ed.gov/ncee/wwc/

Wolf, M., & Denckla, M. B. (2005). *Rapid Automatized Naming and Rapid Alternating Stimulus Tests*. Austin, TX: PRO-ED.

Woodcock, R. W. (1998). *Woodcock Reading Masters Tests–Revised examiner's manual*. Circle Pines, MN: American Guidance Service.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, *44*, 41–61.