Reading and Writing (2020) 33:2697–2724
https://doi.org/10.1007/s11145-020-10060-2

Effects of a read aloud intervention on first grade student vocabulary, listening comprehension, and language proficiency

Doris Luft Baker[1,3] · Lana Santoro[2] · Gina Biancarosa[2] · Scott K. Baker[1,2] · Hank Fien[2] · Janet Otterstedt[2]

Published online: 6 June 2020

Abstract
We examine the effects of a read aloud replication intervention designed to improve the vocabulary, comprehension, and expository and narrative language outcomes of first grade students. Thirty-nine first-grade classrooms from 12 schools were randomly assigned to a treatment (n = 19) or comparison condition (n = 20). Teachers in the treatment condition implemented a 19-week set of read aloud lessons during whole-class read aloud time. Read alouds included the systematic use of narrative and expository texts, before-, during-, and after-reading components, the use of teacher-facilitated text-based discourse, and explicit comprehension instruction. Results indicated main effects of treatment on vocabulary knowledge. Exploratory findings indicated a significant interaction effect of treatment and recommended features of read aloud instruction on all outcomes. Specifically, students of teachers in the treatment condition who were rated higher on adhering to recommended features of read aloud instruction had better outcomes on vocabulary, comprehension, and language outcomes on expository and narrative text than treatment teachers who closely followed intervention materials without dynamically adjusting to student responses. We discuss these findings in the context of other read aloud studies, including a previous study that used the same intervention in a different setting and with a less diverse sample of students.

Keywords Read alouds · Text-based comprehension instruction · Narrative text · Expository text

✉ Doris Luft Baker
dluftdebaker@smu.edu

[1] Department of Teaching and Learning, Simmons School of Education, Southern Methodist University, Dallas, TX 75275, USA

[2] University of Oregon, Eugene, OR 97403, USA

[3] University of Texas, Austin, TX 78712, USA

Introduction

This article presents the results of a replication study of a read aloud intervention in grade 1. In the original study, students in classrooms randomly assigned to the read aloud treatment condition outperformed students in comparison classrooms on two outcome measures: vocabulary and narrative retell (Baker et al., 2013). On two outcome measures, expository retell and a standardized measure of listening comprehension, differences were not significant. The same intervention implemented in the original study was implemented in this replication. The original study was conducted in the Pacific Northwest. The replication was conducted in the Mid-Atlantic region of the U.S.

The value of replication research in education is increasing, both because replications are crucial in scientific research and because they lead to stronger and more accurate policy and practice recommendations (Makel & Plucker, 2014). However, in a study on replication rates, less than 1% (0.13%) of published studies in top education journals were replication studies, likely due to various types of biases (e.g., submission, funding, editor/reviewer, publication, promotion; Makel & Plucker). In education intervention research, replication studies are sought to examine variations in the settings in which studies are conducted, the populations of students being investigated, and the procedures used in training and implementation of treatment and comparison condition practices (Coyne, Cook, & Therrien, 2016; Makel & Plucker, 2014; Travers, Cook, Therrien, & Coyne, 2016).

This replication was undertaken with several of these considerations in mind. First, a larger sample of teachers and students participated. Over three times as many classrooms and students participated in the replication. Effect sizes on statistically significant student outcomes in the original study were moderate (0.42) or large (0.93) and the difference on expository retell though smaller in magnitude (effect size = 0.28) was close to statistically significant ($p = .07$). A larger sample with more power to detect effects could potentially replicate the positive effects and detect a significant effect on expository retell, and perhaps listening comprehension (effect size in the original study = 0.16).

Second, the student sample in the replication study differed in important ways from the original student sample. In the original study, 80% of the students were White; 1% were Black and 13% were Hispanic. In the replication only 24% of students were White. Also, 44% of students in the replication were English learners and more than 100 different primary languages were spoken at home. Third, in the replication the extent to which both treatment and comparison teachers implemented recommended features of read aloud instruction was measured (August, Artzi, Barr, & Francis, 2018; Coyne, Kame'enui, & Carnine, 2011; Wasik, Bond, & Hindman, 2006). Our objective was to explore the extent to which these recommended features of read aloud instruction could help account for the findings in treatment and comparison classrooms.

Research on read alouds

Read aloud instruction is a common activity in U.S. classrooms in the early elementary grades. Students enjoy the experience, as it offers engaging stories and tends to be infused with animated voices and gestures that are amusing and witty. Read alouds also provide students with opportunities to engage in content that helps build background knowledge and understanding of academic topics (Lennox, 2013; Parsons & Bryant, 2016). Furthermore, learning demands in the Common Core State Standards (CCSS; 2010) and other state-specific standards means students are expected to acquire specific knowledge, vocabulary, and language proficiency skills earlier and in greater depth than in previous decades. Read alouds, in addition to being enjoyable, can be a mechanism for students to acquire knowledge, make sense of complex content, and develop discourse skills on specific topics. For example, posing and answering inferential questions about text can help students actively participate in the read aloud lesson. Connecting read aloud topics to other academic topics such as science or social studies can help students acquire discipline-specific knowledge and see connections across topics. Connecting read aloud events to students' personal experiences in and out of school settings can help students connect read aloud content to their own lives (Santoro, Baker, Fien, Smith, & Chard, 2016; Giroir, Grimaldo, Vaughn, & Roberts, 2015; Wasik et al., 2006).

The number of studies on the impact of read aloud practices on student learning outcomes is not large, but the studies that have been conducted demonstrate positive effects in early childhood and elementary school settings. Three meta-analyses have summarized these findings (i.e., National Early Literacy Panel, 2008; Swanson et al., 2011; What Works Clearinghouse, 2007). The National Early Literacy Panel (NELP, 2008) examined the impact of 19 experimental and quasi-experimental studies on shared storybook reading interventions published between 1985 and 2003. Studies were conducted with children from birth to age 5 in home- and center-based settings. Shared storybook reading is a type of read aloud practice where parents or teachers read aloud to children individually or in small or large groups. Before, during, and after the read aloud, the adult facilitates interactive discussions with children about the text. Moderate impacts were found on oral language (effect size = 0.57) and print awareness (effect size = 0.50).

The What Works Clearinghouse (WWC, 2007) reviewed eight read aloud intervention studies (four not included in the NELP review) conducted with children age 3–5. Three studies addressed shared storybook reading and outcomes were rated as potentially positive for early reading and writing (effect size = 0.70) and mixed for oral language (effect size = 0.08). There were no discernible effects for print knowledge (effect size = 0.10). Five studies focused on dialogic reading, which the WWC defined as the use of a specified prompting or cueing system to promote student discourse and comprehension during read aloud. Dialogic reading also includes the use of role-playing and group discussion after reading where the child might play the role of the storyteller with adult support. The overlap between dialogic reading and interactive storybook reading is considerable. The essential purpose of each is to get children to participate verbally and actively in the read aloud experience with an adult. In the five studies, dialogic reading was used individually with students or in small groups. Overall, dialogic reading had a positive effect on oral language (effect size = 0.50) and no discernible effect on phonological processing (effect size = 0.22).

The WWC also reviewed read aloud interventions in kindergarten through grade 3 (K–3) [Institute of Education Sciences (IES) 2007]. The WWC selected commercially available programs for review, not specific interventions of the type reviewed in the early childhood set. One K–3 study met evidence standards for research design quality and was examined for effectiveness. Phillips and colleagues (Phillips, Norris, Mason, & Kerr 1990) investigated the impact of read alouds in the context of little books, texts with high frequency words, simple sentences, and thematic topics. The intervention was delivered to kindergarten children in home and school settings and had a potentially positive effect on general reading achievement (effect size = 0.31; IES, 2007).

Swanson et al. (2011) investigated read aloud interventions delivered in schools (i.e., not by parents at home) to students at risk for learning disabilities. School-based delivery is aligned with the current study and is important for additional reasons. First, teachers provide the read aloud instruction, not a combination of parents and teachers. Second, and most importantly, read aloud instruction in school settings most commonly occurs in whole-group settings, not one-on-one with students or in small groups. In whole-group contexts, teachers may have to employ additional strategies to keep students engaged, actively participate, and derive meaningful benefit.

Swanson et al. (2011) reviewed 29 studies. Ten studies implemented dialogic reading. Other intervention formats included e-books, word elaboration, extended word instruction, music or story-telling programs, text-talk, repeated story book reading, shared book reading, and story reading with limited questioning before, during, and after reading. Findings overall indicated a small effect on oral language (effect size= 0.29), and large effects on phonological awareness (effect size= 0.78), print concepts (effect size= 0.86), vocabulary (effect size= 1.02), and comprehension (effect size= 0.70). Importantly, only two of the 21 treatment–comparison studies (both in preschool settings) met three design issues that substantively strengthen study quality: the use of random assignment to condition, the inclusion of fidelity of implementation procedures, and the use of standardized dependent measures (Raudenbush, Bryk, Cheong, & Congdon, 2004; Shadish, Cook, & Campbell, 2002; U.S. Department of Education, 2003). The current study incorporated these research design features.

In summary, research on read aloud instruction has been conducted primarily in early childhood settings. A smaller number of studies has been conducted in elemary settings. Most research has occurred in home, center, or school-based settings in one-on-one interactions involving an adult and child or in small group formats. Fewer studies have been conducted in whole classroom settings, which is the focus of the current study. Most outcomes have addressed oral language and print awareness, and while some conclusions indicate mixed effects (positive and neutral) most findings indicate moderate, positive effects. Fewer outcomes have been investigated on other aspects of literacy, but the studies conducted have produced positive findings. Relevant to the current study, positive findings have been observed for vocabulary and comprehension.

Features of read aloud instruction

Converging evidence suggests that activities before, during, and after a read aloud lesson can extend student knowledge of content, their understanding of how text is structured to convey information, and improve their vocabulary knowledge and overall language proficiency (August et al., 2018; Baker et al., 2013; Collins, 2016; Lennox, 2013; Neugebauer, Coyne, McCoach, & Ware, 2017; Silverman, Crandell, & Carlis, 2013; Wasik et al., 2006). For example, August et al. (2018) found that second grade students who received extended vocabulary instruction as part of read alouds significantly increased their depth of vocabulary knowledge compared to students who received typical read aloud instruction with target words inserted in the text. Similarly, Silverman et al. (2013) found that 15 min of extended vocabulary instruction significantly improved the vocabulary knowledge of preschool children, compared to typical read aloud instruction where target words were not defined nor discussed in depth. Finally, Wasik et al. (2006) found that activities before and after read alouds that encouraged students to discuss target words and how these words were used in the read aloud books significantly increased students receptive and expressive vocabulary. The approach also improved their comprehension of the read aloud.

In addition to providing in-depth vocabulary supports before, during, and after read alouds, active engagement with complex text can also foster student understanding. Active engagement can include activities such as teachers and students discussing what they know about a topic before reading the text, teachers helping students make connections to other read alouds or to experiences in their own lives, and teachers asking inferential questions that can lead to meaningful discussions about the text (Santoro et al., 2016; Collins, 2016; Giroir et al., 2015; Parsons & Bryant, 2016). In summary, strategically balanced teacher–student interactions before, during, and after reading the text appear to be important to improving student understanding of what they hear and build their competence and confidence in forming ideas about book content that they can then explain to others (Baker et al., 2013; Beck & McKeown, 2007).

The strategies used in the current read aloud intervention incorporated engaging activities before, during, and after read alouds to foster active participation and comprehension. Five strategies for improving comprehension recommended by the National Reading Panel (NRP, 2000) anchored efforts to increase engagement and comprehension: (a) summarizing texts, (b) asking and creating questions, (c) working collaboratively with others, (d) representing texts structurally and graphically, and (e) monitoring comprehension.

Also woven into the structure of the intervention were six evidence-based principles associated with effective instruction recommended by Coyne et al. (2011). First, the core components or big ideas in the domain are highlighted and drive instruction (Coyne et al., 2011). Big ideas in comprehending text include identifying text features, understanding the vocabulary and how specific words are being used, and applying cognitive strategies to determine meaning. Second, students are taught strategies conspicuously to help make learning content, especially abstract content, clear and concrete. Steps are outlined, activities are explained, and tools, such as visual and graphic organizers, are provided. Third, scaffolds are provided to students to help mediate learning. Substantial support from the teacher occurs during initial learning but as students progress these supports are purposefully faded. Fourth, material is integrated strategically to help students make connections between content they have learned and new content they are learning. Fifth, teachers determine, and teach when necessary, the general background knowledge students must possess to learn and acquire new knowledge. Sixth, content is reviewed sequentially, adequately, and cumulatively to help students learn content deeply and relate what they have learned to other content. The read aloud intervention in this study required teachers to apply these specific and general strategies to engage students in productive learning interactions targeting specific academic topics such as learning key characteristics of mammals and reptiles.

Despite recommendations regarding the use of these types of specific and general features of instruction, they are not typically measured in most intervention studies (Pianta & Hamre, 2009), including in read aloud interventions. Consequently, we do not know the degree to which these recommended approachers are occurring or their association with student outcomes. In this study we measured recommended features of read aloud instruction in both treatment and comparison classrooms to explore their potential impact on student outcomes.

We also measured treatment fidelity in treatment classrooms to determine if all the components of the read aloud lesson were implemented (Harn, Parisi, & Stoolmiller, 2013). For example, treatment fidelity for Lesson 1 included the following items: (a) Teacher sets purpose for reading by telling students they are starting a new book; (b) Teacher tells/guides students to make text-to-text connections (e.g., last book was about sea turtles, this book is about a land turtle); (c) Teacher guides students to discuss the first thing you do with a new book (e.g., identify the purpose for reading by asking, “Is this an information book or a story book?”). Examples of recommended features of instruction and treatment fidelity forms for expository and narrative lessons are available from the first author.

Two objectives were pursued in this study: (a) estimate the effects of a read aloud intervention on student outcomes in an attempt to replicate the findings from a previous study (Baker et al., 2013); and (b) explore whether recommended features of read aloud instruction were associated with student outcomes in both treatment and comparison classrooms.

We hypothesized that the effects of the read aloud intervention would be replicated. That is, we expected to observe effects on vocabulary knowledge, and narrative and expository retells. We also expected that greater power in the replication might result in an effect on listening comprehension. Regarding the use of recommended read aloud practices, we expected treatment classrooms to implement more recommended practices than comparison classrooms, given that the intervention design attempted to directly and indirectly account for these features. We hypothesized that there would be an association between the use of recommended practices and student outcomes.

Method

Participants

Blocking on school, 39 first-grade classrooms in 12 schools in the Mid-Atlantic region were randomly assigned to a treatment or comparison read aloud condition. Nineteen classrooms were in the comparison condition, and 20 classrooms were in the treatment condition. Ten schools were located in urban settings and two schools were located in rural settings. Nine of the 12 schools were schoolwide Title 1 schools, and 29 of the 39 classrooms were in these schools.

Teachers

All 39 classroom teachers were female, and 36 were White. Two teachers were Pacific Islanders and one was Hispanic. Their mean age was 34, ranging from 21 to more than 55 years old. Teachers had an average of 9.4 years of teaching experience, ranging from 1 to 31 years. Regarding education background, 54% of the teachers had a bachelor's degree, and 39% had a master's degree or higher; 71% had a specialization in elementary education or in elementary education and early childhood. Differences in teacher demographics between treatment and comparison conditions were not significant.

Students

A total of 638 students participated in the study, 317 in treatment classrooms and 321 in comparison classrooms. Forty-three percent of students were female; 24% were White; 22% were Black; 29% were Hispanic; 18% were Asian; and 6% were multiracial. Eighteen percent received English as a Second Language services, and 14% received special education services. In the district, 47% received free or reduced lunch prices. English learners were 44% of the student sample. By school, English learners ranged from 20 to 60% of the student population.

Treatment condition

In both the original study and replication, the following features guided the implementation of the read aloud intervention in the treatment condition. Read aloud instruction in treatment classrooms consisted of 24 books, 12 narrative and 12 expository. Books were selected taking into account the following criteria: relevance of the topic for first graders, book length, cost, availability (in libraries or for purchase), text coherence (e.g., a beginning–middle–end structure in narrative texts; e.g., basic features of mammals highlighted in information texts), alignment of text with state science standards, and diversity. In terms of diversity, we selected texts to reflect both male and female characters, different cultures and ethnicity groups, and different settings and geographical locations. Some of these features represented the overall quality of the text and fit within typical read aloud lessons (e.g., relevance, coherence, diversity, standards alignment, length) and others represented replicability in other studies and dissemination efforts (e.g., cost, availability).

Two books, an expository text and a narrative text, were part of a thematic unit that lasted approximately 2 weeks. The three themes in the curriculum focused on animals: mammals, reptiles, and insects. Each unit included six or seven lessons. Three lessons focused on the expository text, and three or four lessons (depending on the unit) focused on the narrative text. Each lesson lasted about 30 min. For the insects theme, for example, the first unit focused on the general animal category (insects) and included an information book about insects and a narrative book featuring many different kinds of insects as story characters. The following units in the insects theme contained specific examples from the general animal category, in this case butterflies and ladybugs. A teacher's guide provided step-by-step guidance on the types of activities and questions teachers should engage in before, during, and after reading the text with students. The entire read aloud intervention lasted 19 weeks.

Teacher implementation of read aloud lessons followed principles of explicit instruction, which was laid out for treatment teachers in a very detailed implementation guide. A consistent lesson framework entailed teachers demonstrating read aloud practices (model), teachers and students working together on these practices (lead), and students engaging in these practices on their own (independent practice; Santoro, Chard, Howard, & Baker, 2008). Instruction incorporated specific features of effective instruction described by Coyne et al. (2011). *Before text reading*, teachers lead students in identifying the book type (e.g., expository or narrative) and made predictions about what the text might be about. Teachers also provided definitions of, and practice with, critical vocabulary to build background knowledge and student understanding of the content.

During text reading teachers instructed students in how to comprehend text, such as finding details in the text that would help them draw reasonable inferences. For example, when learning about critical features of reptiles, students looked for details in the text such as "cold-blooded," "scales and plates," and "hatch from an egg" to help determine whether turtles are reptiles. During text reading, teachers also taught words that were new or difficult in meaning as they occurred in the text.

After text reading teachers modeled a narrative or expository retell using a common framework. Students then practiced retelling the text using this framework (Santoro et al., 2016). With a narrative text, one example of a framework was a visual organizer that included icons for the main character and three questions that students were taught to include in their retells: "What happened first?", "What happened next?", and "What happened at the end?". For expository text retells, students answered the following questions that are typical in a K–W–L chart: "What did you think you knew? What did you want to know? What did you learn?" Over time, students retold texts with no teacher model at the beginning. All student activities were practiced in pairs or sometimes in small groups. Throughout the intervention, the idea was to use text-based discourse to stimulate student academic language use, and to prompt student vocabulary use and language-based elaborations. For more detailed information about the intervention, see (Baker et al., 2013; Santoro et al., 2008, 2016).

Training treatment teachers

Training procedures for treatment teachers were accomplished in one full day of training, which was used used in the original study and occurred prior to implementation. Topics covered research supporting the read aloud approach, as well as a detailed summary and overview of the lessons. Teachers practiced implementation by modeling a lesson, and video clips were used for teacher discussions of features of instruction and implementation. Given that an important component of read aloud instruction was dialogic interactions between teachers and students, and among students, teachers were trained on how to have students work in dyads on prescribed comprehension tasks, such as retells.

In the replication, all teachers received three additional hours of training that addressed instruction specifically with English learners. This did not occur in the original study. This was provided in the replication in response to coaching visits with all intervention teachers that occurred during Week 3 of the intervention. The three additional hours of training occurred during Week 4. Members of the research team discussed how to make adaptations to the program to support English learners. These adaptations reflected research-based practices for teaching English learners such as using additional repetitions when presenting word definitions, emphasizing sections of expository text that contained critical information, providing additional visual supports, using sentence frames consistently, and allowing newcomers to provide one-word answers until they were more confident speaking English (Baker, Al Otaiba, Ortiz, Correa, & Cole, 2014).

Once the intervention began, a staff member with read aloud expertise observed each teacher early in the intervention during her read aloud instruction. The staff member then met with the teacher to provide feedback on the content and delivery of instruction. During Week 9 (about half way through the intervention), a staff member provided a follow-up half-day training to treatment teachers to review lesson components and present details of the remaining lessons in the program. This training also occurred in the original study.

Comparison condition

Comparison teachers were given the same books as the treatment teachers. They were encouraged to use these texts as much (or as little) as they wanted and in whichever way they believed would be most beneficial to comprehension development. For evaluation purposes, one-half of the comparison teachers were required to follow specific implementation procedures during Week 8. The other half of the comparison teachers followed the same procedures during Week 14. Assignment to week was random. During their assigned week, comparison teachers used the specific books we identified each day that week and they taught a read aloud lesson in a manner they believed would be beneficial to students. Each day's lesson was to last for about 30 min. Based on direct observations, none of the comparison classrooms had access to or used the implementation guide that treatment teachers used to provide read aloud instruction.

Student measures

To assess the impact of read aloud instruction on student outcomes, we assessed students on listening comprehension, vocabulary knowledge, and expository and narrative text retells.

Gates-MacGinitie test of reading comprehension, listening comprehension subtest (MacGinitie, MacGinitie, Maria, Dryer, & Hughes, 2000)

The listening comprehension subtest of the Gates-MacGinitie was administered at pretest and posttest to evaluate listening comprehension. During administration, the examiner read a short story to students, repeated a short segment from the story, and then prompted students to select one of three pictures that went with that part of the story (MacGinitie et al., 2000, p. 96). Reliability is reported as 0.81 for the fall of first grade (Kuder-Richardson 20 coefficient; MacGinitie et al., 2000). The average correlation of the Listening Comprehension subtest with three other reading subtests administered concurrently in the fall of first grade is reported as 0.55 (MacGinitie et al., 2000). Predictive validity of the Gates-MacGinitie for the fall and spring of first grade is reported as 0.74 (MacGinitie et al., 2000). For study reliability, we double scored 20% of test protocols and achieved 100% accuracy.

Depth of vocabulary knowledge (DOK)

The DOK measure was developed following procedures used by Eller, Pappas and Brown (1988) and further developed in Baker et al. (2013). The measure was individually administered to students at pretest and posttest. Each DOK assessment consisted of 16 words sampled from a pool of 33 narrative-related and 41 expository-related words from texts used in the study and included in district and state curriculum standards for animal science. On the DOK, examiners asked students to define a word and use it in a sentence. Students received one score for defining the word (0–2 scale) and a second score for using the word in a sentence (0–2 scale). Interrater agreement based on total score was 0.95 and internal consistency, as measured by Cronbach's alpha, was 0.80 and 0.87 at pretest and posttest, respectively.

Expository retells: strong narrative assessment procedure (SNAP)

We applied the SNAP administration and scoring procedure to assess student comprehension of expository and narrative text at pretest and posttest. The SNAP (Strong, 1998) is a standardized measure of listening comprehension that was individually administered to all students. For the expository retells, students listened to an audiotape of a text about killer whales (this text was not used in the intervention), after which they were prompted to tell what they remembered. The number of correct concepts was used as an index of comprehension in the analysis. Two raters coded 20% of the protocols and interrater reliability was 0.98. In a study examining first grade expository retells, Moss (1997) found that first grade students were able to include key ideas and details along with cohesive information.

Narrative retells

These retells also used the SNAP administration and scoring procedures developed by Morrow (1985). Students listened to a tape-recorded story as they viewed a wordless picture book. Auditory signals were used to cue page turning. At the end of the story, students retold the story in their own words without the use of the picture book. Story components and plot episodes were counted separately and each was used as an index of comprehension. Two raters coded 20% of the protocols, and interrater reliability for total score was 0.85. Previous research supports retells administered and scored this way with students in kindergarten through second grade to evaluate narrative comprehension (Dougherty & Stahl, 2009; Paris & Paris, 2003).

The expository and narrative retells were transcribed using the Systematic Analysis of Language Transcript (SALT) software (Miller & Chapman, 1993). Rater training involved a process where retells were initially scored as a group, and then independently. During training, group discussions were used to reconcile scoring differences. After training, each rater independently scored a narrative and an expository retell. All raters achieved agreement of at least 0.80 on each retell before coding independently. Interrater agreement was determined by counting the number of line-by-line agreements and line-by-line disagreements, then dividing by the total number of agreements and disagreements. All coders were blind to condition.

Read aloud instruction measures

Treatment fidelity

We used two types of measures to assess aspects of read aloud instruction. The first measure was a treatment fidelity measure designed to assess if teachers in the treatment condition implemented key aspects of the read aloud intervention as intended. We used this measure in treatment classrooms only. This fidelity measure addressed basic issues such as whether teachers used the targeted books during the lesson as well as more complex aspects of instruction such as whether teachers followed detailed suggestions in the teacher's guide, including using a model, lead, test framework, engaging in specific activities *before*, *during*, and *after* text reading, and providing explicit explanations and prompts during the lesson.

Fidelity was coded on a 0–1 scale according to the presence or absence of each component. In this study, fidelity of implementation was 0.73 for lessons using information text, and 0.77 for lessons using narrative text. This index of fidelity is somewhat lower than many fidelity measures associated with intervention implementation in research studies, in part because it was not expected that teachers would necessarily address all lesson components during each lesson. Teachers were expected to address as many components as possible, and not skip major activities (e.g., the before, during, or after reading sections). They were expected to make their own instructional decisions about how to address specific components as the lesson proceeded. Consequently, the fidelity estimates matched our expectations and were in line with what we observed in the original study.

Recommended features of read aloud instruction

The second measure focused on recommended features of read aloud instruction as described in NELP (2008), the NRP (2000), and Coyne et al. (2011). We used this measure in both treatment and comparison classrooms. Features included aspects of instruction such as teachers summarizing texts, asking and generating questions, helping students work cooperatively with each other, and representing texts structurally and graphically. One coding form addressed lessons on expository text and a second form addressed lessons on narrative next.

Read aloud lessons were audio recorded and then coded on a 0–2 rating scale (i.e., 0 = *not done*; 1 = *done*; 2 = *done well*). Distinguishing between *done* and *done well* represented whether a feature was simply present in the lesson or whether the feature was implemented more extensively, with teacher explanations, opportunities for students to practice, the presence of scaffolds to support student discourse and comprehension, and differentiation of instruction based on student need. Each teacher had lessons coded for expository text and for narrative text. Scores were averaged to get an overall score per teacher. Interrater reliability of lessons double coded was 0.95.

Observation training

Six individuals with classroom teaching experience (e.g., retired or substitute teachers) participated in a full day of training focused on coding read aloud lessons. Training was provided on both the fidelity measure and the measure of recommended features of read aloud instruction. During training, coders reviewed all items and definitions on both measures, then coded together and independently audio files of sample lessons. Audio files used in the training were not from classrooms participating in the study. To be certified to code actual study lessons, all coders had to obtain interrater agreement of 0.85 or greater on the treatment fidelity measure and on the recommended features of instruction measure.

Data collection procedures

Read aloud instruction data

During Weeks 8 and 14 treatment and comparison teachers had their read aloud lessons audio recorded and then coded for analysis. Teachers had one narrative lesson and one expository lesson audio recorded. Half of the teachers had their two lessons recorded during Week 8 and the other half were recorded during Week 14. Twenty-five percent of the recorded lessons were double-coded for reliability purposes. Results of interrater reliability calculated as the number of agreements divided by the number of agreements plus disagreements was 0.92 for double-coded lessons.

Student assessments

For student assessments, data collectors received a full day training. Training focused on reviewing and practicing the administration and scoring of each measure. Training also covered procedures for working in schools, communicating with students, including English learners, the use of neutral encouragement during assessments, and standards for mandatory reporting and confidentiality. Before collecting data with students, data collectors had to administer assessments with 100% accuracy, based on procedural checklists aligned with administration protocols, and achieve at least 95% interrater agreement on scoring.

Data analysis procedures

To analyze the data, we used a two-level hierarchical linear model (HLM) with two levels, student, and classroom. We did not include a school level because initial analysis using a three-level model indicated that between school variance was not significant (this analysis is available by request from the authors). For each outcome, a null model (Model 1) was run, followed by the addition of listening comprehension at pretest (Pretest) at Level 1 (Model 2). In the main analysis, the Level 2 predictor, condition assignment (Group; 0 = comparison, 1 = treatment) was added as a predictor of the intercept (Model 3).

In exploratory analyses, the score on recommended features of read aloud instruction was added as a predictor of the intercept (Model 4). Model 5 included an interaction term for condition assignment by the score on recommended features of read aloud instruction (Group × Features) as predictor of intercept (Model 5). Pretest scores were group-mean centered, and Read Aloud Features and the Group × Features interaction were grand-mean centered. The final model (Model 5) tested for each outcome appears below.

Level 1 (Student):

$$Y_{ij} = \beta_{0j} + \beta_{1j}Pretest + r_{ij}$$

Level 2 (Classroom):

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Group + \gamma_{02}Features + \gamma_{03}Group \times Features + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

Combined model:

$$Y_{ij} = \gamma_{00} + \gamma_{01}Group + \gamma_{02}Features + \gamma_{03}Group \times Features + \gamma_{10}Pretest + u_{0j} + r_{ij}$$

In the combined model, Yij represents the outcome for student i in classroom j. γ00 represents the outcome for students with an average pretest score in a comparison classroom with an average score on recommended features of read aloud instruction. γ01 represents the difference in outcomes for students with an average pretest score in an intervention classroom with an average read aloud features score. γ02 represents the difference in outcomes for students with an average pretest score in classrooms with a read aloud features score that is above or below average. γ03 represents the additional difference above or below average on read aloud features in outcomes for students with an average pretest score in treatment classrooms. γ10 represents the difference in outcomes for students with pretest scores above or below their classroom's average in a classroom with an average read aloud features score. Finally, u0j represents the random effect, or residual, associated with classrooms, while rij represents the random effect, or residual, associated with students. We estimated effect sizes by (a) comparing the final model to prior models to calculate the change in pseudo-R^2, and (b) examining the effect of the parameter estimate for being one standard deviation above or below the mean relative to the standard deviation for the model (Raudenbush & Bryk, 2002).

Missing data

Of the 638 participating children, Table 1 shows the number of students who were administered assessments at pretest and posttest by condition. A Chi square test revealed no relation between missingness and condition assignment, $\chi^2(1, N=638)=1.22$, $p=.29$. A series of one-way between subjects ANOVA analyses were conducted to determine whether children missing data on one or more posttest measures differed significantly from those not missing data on any posttest measure. The ANOVAs indicated no significant differences between these groups, with the exception of 46 children who were missing the Vocabulary Total Score measure $F(1, 577)=4.92$, $p=.03$. Students missing at least one posttest result other than the Vocabulary Total Score measure tended to have lower vocabulary scores at posttest. As a result, all subsequent analyses were conducted with all students with available data for a particular outcome, as opposed to limiting the sample to only those children with complete data, which might have resulted in biased estimates.

Of the 39 classrooms, three of the 20 intervention classrooms were missing data on the recommended features of read aloud instruction. Two of these three classrooms had very small numbers of students (classroom $n=5$, 6, and 17) because they were mixed-grade classrooms. Thus, overall student sample size was not dramatically affected. Due to randomization at the classroom level, we conducted a post hoc power analysis to determine effects on power to detect main effects of treatment. We ran a post hoc power analysis for detecting small ($d=0.20$) and medium ($d=0.40$) effects using Optimal Design 3.01 for a two-level cluster-randomized trial. We set cluster size to 17, which was the sample mean, and examined two intraclass correlations (ICCs) of 0.03 and 0.17, which were the observed minimum and maximum ICCs for student outcomes (for listening comprehension and vocabulary respectively). For detecting a medium main effect of treatment, we found that achieved

Table 1 Descriptive statistics for student and classroom variables for full sample of 39 participating classrooms and limited sample of 36 classrooms with complete classroom data

Student variables	Control (19 classrooms)			Intervention (20 classrooms)			Intervention (17 classrooms)			Total (39 classrooms)			Total (36 classrooms)		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Pretest															
Gates	301	12.0	4.8	305	12.1	4.5	277	11.9	4.5	606	12.1	4.7	578	11.9	4.7
Vocabulary depth-of-knowledge	168	9.1	8.4	178	9.1	7.9	164	9.1	8.0	346	9.1	8.1	332	9.1	8.2
Narrative retelling: components	173	5.1	3.1	187	5.7	3.4	170	5.6	3.4	360	5.4	3.3	343	5.3	3.3
Narrative retelling: episodes	173	11.5	8.2	187	11.8	6.8	170	11.6	6.6	360	11.7	7.5	343	11.5	7.4
Posttest															
Gates	297	15.4	4.1	294	15.4	4.0	266	15.3	4.1	591	15.4	4.0	563	15.3	4.1
Vocabulary depth-of-knowledge	287	21.3	12.4	292	26.0	13.4	265	26.5	13.4	579	24.7	13.1	552	23.8	13.2
Narrative retelling: components	280	5.1	1.5	289	5.1	1.5	262	5.2	1.5	569	5.1	1.5	542	5.1	1.5
Narrative retelling: episodes	280	14.4	5.7	289	14.3	5.5	262	14.5	5.4	569	14.4	5.6	542	14.5	5.5
Expository retelling: concepts	290	5.1	3.2	282	5.5	3.7	259	5.5	3.7	572	5.3	3.5	549	5.3	3.5
Classroom variable															
Recommended features	19	0.4	0.1	NA	NA	NA	17	0.7	0.2	NA	NA	NA	36	0.5	0.2
Implementation fidelity	NA	NA	NA	NA	NA	NA	17	0.8	0.1	NA	NA	NA	NA	NA	NA

Gates = Gates-MacGinitie Listening Comprehension Test. NA = not applicable power for 39 clusters ranged from 0.74 to 0.98 for ICCs of 0.17 and 0.03 respectively. Power to detect small main effects of treatment for 39 clusters ranged from 0.26 to 0.53 for ICCs of 0.17 and 0.03 respectively. Once the three classrooms were dropped, achieved power for medium effects ranged from 0.71 to 0.98 for ICCs of 0.17 and 0.03 respectively and for small effects from 0.24 to 0.51. Thus, the power to detect small effects was weak prior to dropping classrooms, but power was not substantially affected by dropping three classrooms.

In addition, we conducted a one-way between subjects ANOVA to examine significant differences among children in the classes missing instructional features data versus those not missing this data. The three classes missing data had significantly different or nearly significantly different means compared to the full sample on the Gates pretest, $F(1, 604) = 9.69$, $p = .002$, Gates posttest, $F(1, 589) = 3.82$, $p = .051$, and major components present in the narrative retelling, $F(1, 567) = 3.90$, $p = .049$. Students in the classrooms with missing data on the recommended features of read aloud instruction tended to have significantly higher listening comprehension at pretest (by about 3 points) and lower narrative retelling scores (by about 0.6 points) at posttest as compared to classrooms not missing features data. As a result, because all classrooms missing recommended features data were intervention classrooms, exploratory analyses incorporating features of instruction may underestimate effects on listening comprehension and over-estimate effects on retelling skills.

Results

Descriptive results and group equivalence

Table 1 presents the means, standard deviations, and sample sizes for student measures and the recommended features of read aloud instruction. Statistics are reported for all comparison classrooms ($n = 19$), and for two groups of treatment classrooms—all classrooms with student data ($n = 20$) and the subset of those classrooms with both student and instruction data ($n = 17$).

On most measures the mean differences are minimal between the full and restricted samples of treatment classrooms. On two significant differences (i.e., Gates-MacGinitie at pretest and Narrative Retelling Components), the difference between the full and restricted treatment classrooms amounted to less than 1 point, suggesting that although the students in these classrooms differed significantly from students in all other classrooms including other treatment classrooms, these three classrooms had little effect on the relevant means for the treatment classrooms. This is most likely due to the very small class sizes in classrooms missing recommended features data.

To examine whether the recommended features data were distinct from implementation fidelity data, we correlated the scores from the implementation fidelity data and the recommended features data across observations in a treatment class (i.e., before, during, and after instruction). Correlations in treatment classrooms were moderate ($r = 0.63$), suggesting that the two coding tools were capturing similar, yet distinct features of read aloud instruction.

Finally, treatment and comparison means at pretest for the full sample (classroom $n = 39$) and restricted sample (classroom $n = 36$) data were compared on pretest measures using a series of one-way between-subjects ANOVAs. Classroom pretest results did not differ significantly on pretest listening comprehension for the full sample, $F(1, 604) = 0.13$, $p = .72$, versus the restricted sample, $F(1, 576) = 0.10$, $p = .75$. Similarly, no significant differences existed at pretest on vocabulary for the full sample, $F(1, 344) = 0.001$, $p = .98$ versus the restricted sample $F(1, 330) = 0.002$, $p = .97$. Likewise, no significant differences were observed for narrative retellings at pretest for both the major components and plot episodes scores: full sample respectively, $F(1, 358) = 2.82$, $p = .09$ and $F(1, 358) = 0.22$, $p = .64$, restricted sample $F(1, 341) = 2.25$, $p = .14$ and $F(1, 344) = 0.04$, $p = .84$, respectively.

Main effect results

Tables 2, 3, 4 and 5 present the results of HLM analyses examining main effects analyses on four outcome variables (i.e., listening comprehension, vocabulary, expository retells and narrative retells-major components). Additional tables are available upon request from the first author. Table 2 presents the results on the Gates-MacGinitie Listening Comprehension measure. Model 1 is the unconditional model and Model 2 shows the influence of the Gates-MacGinitie pretest measure on the Gates-MacGinitie posttest measure. Model 3 is the model of interest and shows that the main effect of condition (the "group" row under fixed effects) on the Gates-MacGinitie outcome was not statistically significant.

The same three types of models are presented in Tables 3, 4 and 5 to show the results on the other outcome measures: Vocabulary (Table 3), expository retelling (Table 4), and major components in narrative retelling (Table 5). Table 3 shows that the effect of condition (treatment or comparison) on vocabulary was statistically significant ($\gamma_{01} = 4.66$, $t = 2.35$, $p = .025$). In other words, students in the treatment condition outperformed students in the comparison condition on the depth of knowledge vocabulary measure. The effect size was 0.40. The effect of condition on the other three outcomes were not statistically significant. Mean scores were virtually identical on all three measures.

Exploratory findings

Observations of recommended read aloud practices

Table 1 also presents descriptive data on recommended features of read aloud instruction in treatment and comparison classrooms. Analysis indicated that treatment classrooms scored significantly higher than comparison classrooms on these features, $F(1, 34) = 24.61$, $p < .001$. This difference is not surprising, given that read aloud treatment instruction was developed in part to align with these recommendations.

Tables 2, 3, 4 and 5 also present the findings of exploratory analyses examining the association between recommended features of read aloud instruction and

Table 2 Fixed effects, random effects, and fit statistics for series of models predicting Gates-MacGinitie Listening Comprehension Raw Scores

Fixed effects	Model 1				Model 2				Model 3				Model 4				Model 5			
	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*
Intercept, γ_{00}	15.3	0.22	71.16	< .001	15.25	0.22	68.29	< .001	15.37	0.31	49.23	< .001	15.51	0.37	42.41	< .001	14.95	0.36	40.99	< .001
Group, γ_{01}									− 0.26	0.45	− 0.57	.57	− 0.56	0.60	− 0.94	0.36	− 0.56	0.53	− 1.07	.29
Features, γ_{02}													1.20	1.53	0.78	0.44	− 3.48	1.95	− 1.79	.08
Group × Features, γ_{03}																	8.92	2.70	3.31	.002
Pretest, γ_{10}					0.63	0.03	22.02	< .001	0.63	0.03	22.02	< .001	0.63	0.03	22.02	< .001	0.63	0.03	22.03	< .001
Random effects	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*
Intercept (Class-level), u_{0j}	0.75	0.56	54.87	.017	1.10	1.22	108.58	< .001	1.12	1.26	108.36	< .001	1.14	1.29	107.03	< .001	0.93	0.87	81.97	< .001
Student-level, r_{0j}	4.03	16.23			2.87	8.25			2.87	8.25			2.87	8.24			2.87	8.24		
Fit statistics																				
Deviance	3047				2711				2712				2707				2695			
u_{0j} pseudo-R^2 from Model 2	–				–				–				–				0.29			

Table 3 Fixed effects, random effects, and fit statistics for series of models predicting depth of vocabulary knowledge

Fixed effects	Model 1				Model 2				Model 3				Model 4				Model 5			
	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*
Intercept, γ_{00}	23.5	1.05	22.3	< .001	23.5	1.05	22.27	< .001	21.3	1.37	15.54	< .001	22.3	1.56	14.28	< .001	20.4	1.63	12.51	< .001
Group, γ_{01}									4.66	1.98	2.35	.025	2.39	2.58	0.93	.36	2.25	2.37	0.95	.35
Features, γ_{02}													8.91	6.58	1.36	.19	− 7.67	8.73	− 0.88	.39
Group × Features, γ_{03}																	32.1	12.1	2.65	.013
Pretest, γ_{10}					1.71	0.09	18.44	< .001	1.71	0.09	18.44	< .001	1.71	0.09	18.44	< .001	1.71	0.09	18.44	< .001
Random effects	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*
Intercept (Class-level), u_{0j}	5.45	29.8	142.6	< .001	5.82	33.8	241.2	< .001	5.41	29.2	204.3	< .001	5.32	28.3	191.9	< .001	4.81	23.1	159.5	< .001
Student-level, r_{0j}	11.9	142.9			9.20	84.6			9.20	84.6			9.2	84.6			9.20	84.6		
Fit statistics																				
Deviance	4163				3909				3899				3893				3878			
u_{0j} pseudo-R^2 from Model 2	–				–				0.14				0.16				0.32			

Table 4 Fixed effects, random effects, and fit statistics for series of models predicting Expository Retelling Scores

Fixed effects	Model 1				Model 2				Model 3				Model 4				Model 5			
	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*
Intercept, γ_{00}	5.25	0.21	25.0	< .001	5.24	0.21	24.97	< .001	5.06	0.29	17.46	< .001	5.50	0.31	17.94	< .001	5.07	0.32	16.09	< .001
Group, γ_{01}									0.38	0.42	0.91	.37	− 0.54	0.51	− 1.06	.30	− 0.56	0.46	− 1.21	.23
Features, γ_{02}													3.57	1.28	2.79	.009	0.06	1.68	0.04	.97
Group × Features, γ_{03}																	6.75	2.33	2.90	.007
Pretest, γ_{10}					0.32	0.03	10.94	< .001	0.32	0.03	10.94	< .001	0.32	0.03	10.94	< .001	0.32	0.03	10.94	< .001
Random effects	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*
Intercept (Class-level), u_{0j}	0.91	0.83	75.30	< .001	0.99	0.97	93.56	< .001	0.99	0.98	91.16	< .001	0.84	0.71	72.82	< .001	0.69	0.47	57.90	.004
Student-level, r_{0j}	3.26	10.64			2.93	8.57			2.93	8.57			2.93	8.58			2.93	8.57		
Fit statistics																				
Deviance	2762				2661				2659				2651				2638			
u_{0j} pseudo-R^2 from Model 2	–				–				–				0.27				0.52			

Table 5 Fixed effects, random effects, and fit statistics for series of models predicting major components present in narrative retelling

Fixed effects	Model 1				Model 2				Model 3				Model 4				Model 5			
	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*	γ	*SE*	*t*	*p*
Intercept, γ_{00}	5.11	0.09	58.18	< .001	5.11	0.09	58.14	< .001	5.08	0.12	41.12	< .001	5.19	0.14	36.81	< .001	5.01	0.15	33.42	< .001
Group, γ_{01}									0.05	0.18	0.30	.77	−0.17	0.23	−0.72	.47	−0.17	0.22	−0.78	.44
Features, γ_{02}													0.86	0.59	1.46	.15	−0.62	0.80	−0.78	.44
Group × Features, γ_{03}																	2.81	1.10	2.56	.015
Pretest, γ_{10}					0.07	0.01	4.87	< .001	0.07	0.01	4.87	< .001	0.07	0.01	4.87	< .001	0.07	0.01	4.88	< .001
Random effects	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*	*SD*	VC	χ^2	*p*
Intercept (Class-level), u_{0j}	0.36	0.13	66.20	.001	0.37	0.14	69.31	< .001	0.38	0.14	69.00	< .001	0.36	0.13	64.44	< .001	0.31	0.10	54.36	.008
Student-level, r_{0j}	1.44	2.09			1.41	1.99			1.41	1.99			1.41	1.99			1.41	1.99		
Fit statistics																				
Deviance	1879				1861				1864				1859				1853			
u_{0j} pseudo-R^2 from Model 2	–				–				–				0.07				0.29			

outcomes. In each table, Model 4 shows the association between recommended features of read aloud instruction and student outcomes. Model 5 shows whether the interaction effect between treatment condition and recommended features of read aloud instruction has an influence on student outcomes. For all five outcomes (i.e., listening comprehension, vocabulary, expressive retell, narrative retell, and narrative plot episodes), the interaction between read aloud condition and recommended features of read aloud instruction (Model 5) has a statistically significant effect on student outcomes.

A visual depiction of the significant interaction effect of the intervention and listening comprehension, vocabulary, and expository retells is shown in Figs. 1 and 2. Additional figures can be provided upon request to the first author. Students in treatment classrooms where instruction was one standard deviation above the mean on recommended features of read aloud instruction scored significantly higher than (a) students in treatment classrooms where instruction was one standard deviation below the mean, and (b) students in comparison classrooms with instructional features one standard deviation above the mean. This pattern was the same and statistically significant on all five outcome measures. In contrast, within comparison classrooms, students in classrooms where recommended read aloud instructional features were one standard deviation above the mean either scored the same as or lower than students in comparison classrooms below this level. In treatment classrooms, the magnitude of the association, expressed as an effect size, ranged from 0.33 to 0.47. In sum, a higher presence of recommended read aloud instructional features was associated with positive outcomes in treatment

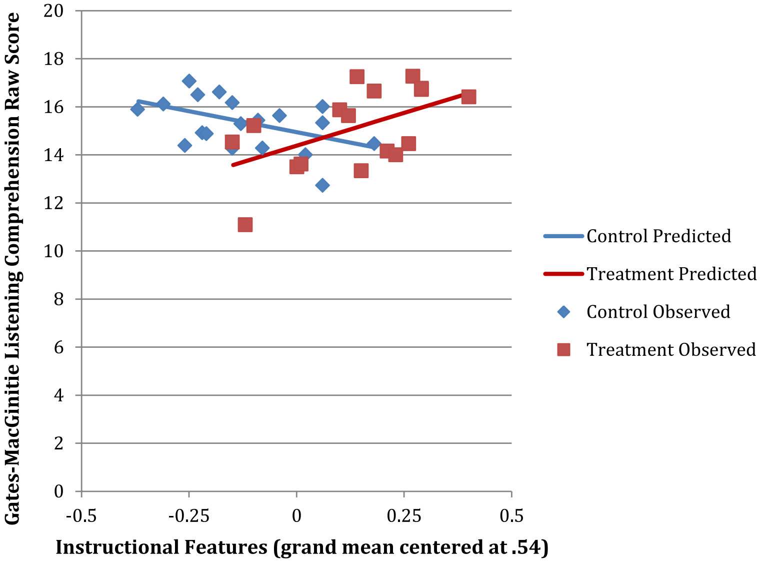

Fig. 1 Effect of interaction between intervention and read aloud instructional features on listening comprehension

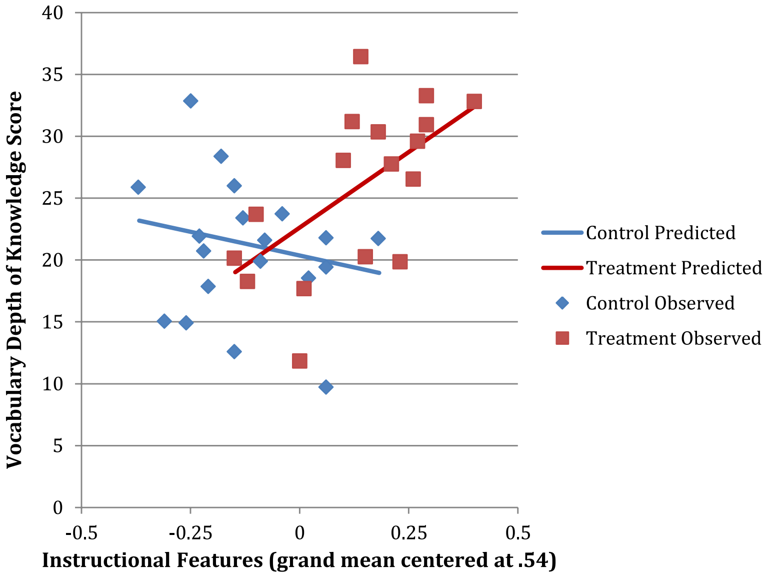

Fig. 2 Effect of interaction between intervention and read aloud instructional features on vocabulary depth of knowledge

classrooms, but the same was not true in comparison classrooms. However, this analysis is exploratory, and the findings should be interpreted cautiously.

Discussion

The main purpose of this study was to examine the effects of a first-grade, read aloud intervention used in whole-group classroom settings serving a diverse population of students. A secondary purpose was to explore differences in the use of recommended read aloud practices in treatment and comparison classrooms, and whether the use of recommended practices correlated with student outcomes on language, vocabulary, and comprehension measures. Results indicated a significant main effect on vocabulary knowledge favoring the treatment group, but no significant effects on the other outcome measures. Two exploratory findings were observed. First, treatment classrooms implemented more recommended features of read aloud instruction than comparison classrooms. Second, in treatment classrooms as the use of recommended features increased from below average (0.25 SDs below the mean) to above average (0.25 SDs above the mean) there was a corresponding statistically significant increase on all five student outcome measures (i.e., listening comprehension, vocabulary, expository and narrative retells, and plot episodes). In comparison classrooms, there was no similar association pattern on any of the student outcome measures. We discuss these findings in the context of similar experimental and quasi-experimental studies, and suggest that recommended features of read aloud instruction, in addition to treatment fidelity, should be measured across conditions when testing the effects of interventions.

Main effects of the read aloud intervention

Results indicate a main effect of the read aloud intervention on vocabulary. This finding replicates the previous study showing a read aloud impact on student vocabulary knowledge (Baker et al., 2013). Our findings also corroborate findings from other studies that have examined the effects of read alouds on vocabulary breadth and depth. These studies included either extended vocabulary instruction outside the read aloud time, or embedded vocabulary instruction within the read aloud time (see, for example, August et al., 2018; Silverman et al., 2013). Findings from August et al. and Silverman et al. indicated that students in the treatment condition learned more target words, and at deeper levels, compared to students in the control condition who received typical read aloud instruction without extended or embedded vocabulary instruction. Effect sizes were moderate to large.

In our read aloud intervention, before text reading, teachers introduced new vocabulary that helped students build background knowledge to be able to understand the content. For example, in a unit about reptiles, teachers introduced characteristics of reptiles (e.g., cold-blooded, they have scales and plates, and they hatch from an egg) before reading the text. After text reading students engaged in conversations focused on the text where teachers encouraged students to use the target vocabulary they had learned in that lesson (e.g., Teachers would say: *I liked what you said about turtles. Now say why turtles are reptiles: because they are cold-blooded, they have scales and plates, and they hatch from an egg).* Other activities used to reinforce key concepts included drawing pictures, writing the new vocabulary words, comparing reptiles to other animals students had learned before, and using the target words to describe different types of animals (see Santoro et al., 2016).

Recommended instructional features

While there was a significant treatment–comparison group effect on vocabulary, differences on the other outcome measures were not statistically significant. We did find, however, that assignment explained 42% of the variance in recommended features of read aloud instruction, and that intervention classrooms had, on average, significantly higher levels of recommended read aloud instructional features than comparison classrooms, ($B = 0.253$, $SE = 0.051$, $p < .001$). It may be that an evidence-based intervention improves the use of recommended instructional features, as suggested by Davis, Palincsar, Smith, Arias, and Kademian (2017), but this instructional effect may not be observed on all types of relevant student impact measures.

The interaction effect between recommended read aloud instructional features and condition suggests that the effects of the read aloud intervention might depend on how teachers delivered the instruction. Our observation measure of recommended features of read aloud instruction was designed to capture key features of read alouds that would be apparent in typical first-grade classrooms. This type of measure, which in this study was used in both treatment and comparison classrooms, has the potential to provide valuable information on how specific interventions, in addition to more general read aloud practices, produce their effects. As Connor et al. (2014) noted, few reading studies include an observation of recommended features of instruction in both conditions. Future research should examine more closely how these features mediate the effect of an evidence-based read aloud intervention on student outcomes.

Replication in context

In the initial read aloud study, the significant main effects resulted in standardized effect sizes of 0.93 on vocabulary and of 0.42 on narrative retell. In the current study, the significant vocabulary main effect resulted in an effect size of 0.40; the main effect on narrative retell was not significant. When comparing this replication to the original study, two reasons may have contributed to differences in effect sizes. First, the larger sample size in the replication may have resulted in more stable impact estimates. Although other read aloud studies (e.g., Swanson, et al., 2011) found large vocabulary effects and moderate to large comprehension effects, it is not clear the extent to which impact estimates may have varied by the size of the instructional group. In the current study, whole group formats were used for instruction.

Second, compared to the original study, this replication included a more diverse sample of students, including a higher percentage of students who were English learners. The more diverse student sample may have tempered effects. In a replication study by Vaughn et al. (2006), which involved an intervention to improve outcomes for first-grade English learners with learning disabilities, impacts were smaller in the replication study compared to the initial study. However, students with learning disabilities were not a large percentage of the sample in the current study.

By examining the student population more closely, Vaughn et al. (2006) found that students in the replication study had significantly different levels of oral language proficiency, which could have explained the diminished effect. In other words, the differences may have had something to do with the focus on students with learning disabilities, the lower levels of oral language profiency in the replication, or both.

The larger percentage of English learners in the current replication study compared to the original study resulted in treatment teachers receiving 3 h of additional training in the replication, which focused specifically on English learners. Although this may have contributed to the significant vocabulary effect, it did not seem sufficient to influence the other areas assessed. Also, pretest performance was very similar in the original and replication samples, suggesting that language proficiency differences were not responsible for the observed outcome differences.

Limitations

Three limitations in this study are important to consider. First, we recorded only two lessons in treatment and control classrooms. Additional lesson recordings could have provided a more stable estimate of the effect of instructional practice on students outcomes. Cost and minimizing classroom disruptions were the primary reasons for collecting recording data at two timepoints only. Nonetheless, the recorded lessons helped us identify differences in the delivery of the instruction among teachers.

Second, we were unable to collect observation data from three of 39 classrooms. However, two of the three classrooms with missing observation data were very small (i.e., classrooms with missing data had $n=5$, 6, and 17 students) and our analysis suggests the overall findings were likely not affected. A third limitation is that student outcome measures might not have been sensitive enough to capture the effects of the intervention. For instance, listening comprehension as measured by the Gates McGinitie, and retells as measured by SNAP, might not have been sensitive to intervention effects because the test content and formats for collecting student responses were quite dissimilar from student experiences in the intervention. However, the measures did appear to capture the interaction between condition and recommended practices. Given the exploratory nature of the interaction effect, future research should examine more closely how standardized measures can capture more nuanced information related to how different approaches, besides the use of recommended instructional features, affect outcomes.

Implications for practice

Findings from this study suggest that read aloud interventions aligned with recommended read aloud instructional features can be beneficial for students on important outcomes, in particular vocabulary knowledge. More speculative is the possibility that in classrooms that implement evidence-based read aloud interventions, greater use of practices associated with recommended instructional features produce additional benefits for students (Davis et al., 2017). Future investigations on the use of read alouds should continue to examine content and quality of instruction before, during, and after reading, and whether there might be unique instructional designs related to these three phases of instruction that maximize the benefits of read alouds to build student vocabulary knowledge and to foster deeper comprehension.

Acknowledgements We would like to acknowledge the work by Dr. Priti Haria in the coordination of the data collection of the project. The research reported here was supported, in part, by the Institute of Education Sciences, U.S. Department of Education, through Grant No. R305G050216 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Dr. Lana Santoro, Dr. Scott K. Baker, and Dr. Hank Fien receive royalties from the sale of the Read Aloud intervention.

References

August, D., Artzi, L., Barr, C., & Francis, D. (2018). The moderating influence of instructional intensity and word type on the acquisition of academic vocabulary in young English language learners. *Reading and Writing, 31,* 965–989. https://doi.org/10.1007/s11145-018-9821-1.

Baker, D. L., Al Otaiba, S., Ortiz, M. S., Correa, V., & Cole, R. (2014). Vocabulary development and intervention for English Language Learners in the early grades. In J. Benson (Ed.), *Advances in child development and behavior* (Vol. 46, pp. 281–338). San Diego, CA: Elsevier. https://doi.org/10.1016/B978-0-12-800285-8.00010-8.

Baker, S. K., Santoro, L. E., Chard, D., Fien, H., Park, Y., & Ottersted, J. (2013). An evaluation of an explicit read aloud intervention in whole-classroom formats in first grade. *The Elementary School Journal*, *113*(3), 331–358.

Beck, I. L., & McKeown, M. G. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *Elementary School Journal, 107,* 251–271.

Collins, M. (2016). Supporting inferential thinking in preschoolers: Effects of discussion on children's story comprehension. *Early Education and Development, 27,* 932–956. https://doi.org/10.1080/10409289.2016.1170523.

Connor, C., Spencer, M., Day, S., Guiliani, S., Ingebrand, S., McLean, L., et al. (2014). Capturing the complexity: Content, type, and amount of instruction and quality of the classroom learning environment synergistically predict third graders' vocabulary and reading comprehension outcomes. *Journal of Educationl Psychology, 106,* 762–778. https://doi.org/10.1037/a0035921.

Coyne, M., Cook, B., & Therrien, J. (2016). Recommendations for replication research in special education: A framework of systematic conceptual replications. *Remedial and Special education, 37,* 244–254. https://doi.org/10.1177/0741932516648463.

Coyne, M., Kame'enui, E., & Carnine, D. (2011). *Effective teaching strategies that accommodate diverse learners*. Upper Saddle River: Pearson.

Davis, E., Palincsar, A. S., Smith, S., Arias, A. M., & Kademian, S. (2017). Educative curriculum materials: Uptake, impact, and implications for research and design. *Educational Researcher, 46,* 293–304.

Dougherty Stahl, K. A. (2009). Synthesized comprehension instruction in primary classrooms: A story of successes and challenges. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 25,* 334–355.

Eller, R. G., Pappas, C. C., & Brown, E. (1988). The lexical development of kindergarten learning from written context. *Journal of Reading Behavior, 20,* 5–23.

Giroir, S., Grimaldo, L. R., Vaughn, S., & Roberts, G. (2015). Interactive read alouds for English Learners in the elementary grades. *The Reading Teacher, 68,* 639–648. https://doi.org/10.1002/trtr.1354.

Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools? *Exceptional Children, 79,* 181–193. https://doi.org/10.1177/001440291307900204.

Institute of Education Sciences. (2007). What works clearinghouse. http://ies.ed.gov/ncee/wwc/. Accessed 10 Sept 2007.

Lennox, S. (2013). Interactive read-alouds—An avenue for enhancing children's language for thinking and understanding: A review of recent research. *Early Childhood Education Journal, 41,* 381–389.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dryer, L. G., & Hughes, K. E. (2000). *Gates-MacGinitie reading tests*. Rolling Meadows: Riverside.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty replication in the education sciences. *Educational Researcher, 43,* 304–316. https://doi.org/10.3102/0013189X14545513.

Miller, J. F., & Chapman, R. S. (1993). *SALT: Systematic analysis of language transcripts*. Language Analysis Laboratory.

Morrow, L. M. (1985). Retelling stories: A strategy for improving young children's comprehension, concept of story structure, and oral language complexity. *Elementary School Journal, 85,* 647–661.

Moss, B. (1997). A qualitative assessment of first graders' retelling of expository text. *Reading Research and Instruction, 37,* 1–13.

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel.* National Institute for Literacy.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects.* Washington, DC: Authors. http://www.corestandards.org/assets/CCSSI/ELAStandards.pdf.

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 Publication No. 00-4769). National Institute of Child Health and Human Development. http://www.nichd.nih.gov/publications/nrp/smallbook.htm.

Neugebauer, S., Coyne, M., McCoach, B., & Ware, S. (2017). Teaching beyond the intervention: The contribution of teacher language extensions to vocabulary learning in urban kindergarten classrooms. *Reading and Writing: An Interdisciplinary Journal, 30,* 543–567. https://doi.org/10.1007/s11145-016-9689-x.

Paris, A., & Paris, S. (2003). Assessing narrative comprehension in young children. *Reading Research Quarterly, 38,* 36–76. https://doi.org/10.1598/RRQ.38.1.3.

Parsons, A. W., & Bryant, C. L. (2016). Deepening kindergarteners science vocabulary: A design study. *The Journal of Educational Research, 109,* 375–390. https://doi.org/10.1080/00220671.2014.968913.

Phillips, L., Norris, S., Mason, J. M., & Kerr, B. (1990). Effect of early literacy intervention on kindergarten achievement. Technical Report No. 520. Champaign: Center for the Study of Reading, University of Illinois at Urbana-Champaign.

Pianta, R., & Hamre, B. (2009). Conceptualziation, measurement, and improvement of classroom prcoesses: Standardized observation can leverage capacity. *Educational Researcher, 38,* 109–119. https://doi.org/10.3102/0013189X09332374.

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood: Scientific Software Inc.

Santoro, L. E., Baker, S. K., Fien, H., Smith, J. L., & Chard, D. (2016). Using read-alouds to help struggling readers access and comprehend complex informational text. *Teaching Exceptional Children*, *48*(6), 282–292. https://doi.org/10.1177/0040059916650634.

Santoro, L. E., Chard, D. J., Howard, L., & Baker, S. K. (2008). Making the very most of classroom read alouds to promote comprehension and vocabulary. *Reading Teacher*, *61*, 396–408.

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

Silverman, R. D., Crandell, J. D., & Carlis, L. (2013). Read alouds and beyond: The effects of read aloud extension activities on vocabulary in Head Start classrooms. *Early Education and Development, 24,* 98–122. https://doi.org/10.1080/10409289.2011.649679.

Strong, C. J. (1998). *The strong narrative assessment procedure (SNAP)*. Eau Claire: Thinking Publications.

Swanson, E., Wanzek, J., Petscher, Y., Vaughn, S., Heckert, J., Cavanaugh, C., et al. (2011). A synthesis of read-aloud interventions on early reading outcomes among preschool through third graders at risk for reading difficulties. *Journal of Learning Disabilities, 44,* 258–275. https://doi.org/10.1177/0022219410378444.

Travers, J., Cook, B., Therrien, W., & Coyne, M. (2016). Replication research and special education. *Remedial and Special education, 37,* 195–204. https://doi.org/10.1177/0741932516648462.

U.S. Department of Education. (2003). Random assignment in program evaluation and intervention research: Questions and answers. Retrieved from http://www.ed.gov/rschstat/eval/resources/randomqa.pdf.

Vaughn, S., Linan-Thompson, S., Mathes, P., Cirino, P. T., Carlson, C., Pollard-Durodola, S. D., et al. (2006). Effectiveness of a Spanish intervention and an English intervention for English language learners at risk for reading problems. *American Educational Research Journal, 43,* 449–487. https://doi.org/10.3102/00028312043033449.

Wasik, B., Bond, M. A., & Hindman, A. (2006). The effects of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology, 98,* 63–74. https://doi.org/10.1037/0022-0663.98.1.63.

What Works Clearinghouse. (2007). *Procedures and standards handbook, version* 3.0. http://www.ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2standards_handbook.pdf.