JSLHR

Research Article

Embedded Instruction Improves Vocabulary Learning During Automated Storybook Reading Among High-Risk Preschoolers

Howard Goldstein,[a] Elizabeth Kelley,[b] Charles Greenwood,[c] Luke McCune,[c] Judith Carta,[c] Jane Atwater,[c] Gabriela Guerrero,[c] Tanya McCarthy,[d] Naomi Schneider,[d] and Trina Spencer[e]

Purpose: We investigated a small-group intervention designed to teach vocabulary and comprehension skills to preschoolers who were at risk for language and reading disabilities. These language skills are important and reliable predictors of later academic achievement.
Method: Preschoolers heard prerecorded stories 3 times per week over the course of a school year. A cluster randomized design was used to evaluate the effects of hearing storybooks with and without embedded vocabulary and comprehension lessons. A total of 32 classrooms were randomly assigned to experimental and comparison conditions. Approximately 6 children per classroom demonstrating low vocabulary knowledge, totaling 195 children, were enrolled.
Results: Preschoolers in the comparison condition did not learn novel, challenging vocabulary words to which they were exposed in story contexts, whereas preschoolers receiving embedded lessons demonstrated significant learning gains, although vocabulary learning diminished over the course of the school year. Modest gains in comprehension skills did not differ between the two groups.
Conclusion: The Story Friends curriculum appears to be highly feasible for delivery in early childhood educational settings and effective at teaching challenging vocabulary to high-risk preschoolers.

Oral language skills have been established as an important and reliable predictor of later reading achievement (National Early Literacy Panel, 2008). Early language proficiency is correlated with reading ability, in particular with reading comprehension (Storch & Whitehurst, 2002). Many children begin school with limited language ability, placing them at high risk of diagnosis with reading disability and with later reading failure (Bishop & Adams, 1990; Catts, Fey, Tomblin, & Zhang, 2002). Many of these children are from families with low income and begin school with language deficits relative to middle-class peers (Hoff, 2013). For example, 33% of children enrolled in Head Start were found to have language delays, with scores more than 1 *SD* below the mean on norm-referenced measures (Nelson, Welsh, Trup, & Greenberg, 2011). Within the domain of language, vocabulary in particular seems to be related to socioeconomic status: Children from low-income families tend to have vocabularies that are much smaller than children from middle-income families (Hart & Risley, 1995; Qi, Kaiser, Milan, & Hancock, 2006). These early language deficits have long-term effects on academic achievement (Walker, Greenwood, Hart, & Carta, 1994).

Early childhood education presents an opportunity to address these language deficits. Preschool experiences with vocabulary predict later reading comprehension (Dickinson & Porche, 2011). It is unfortunate that improving the language experiences of early childhood settings can be challenging (Dickinson, 2011). Given the wide range of language skills of preschoolers, it is critical to adapt instruction to the needs of all children. Multitiered system of supports (MTSS) is an increasingly prevalent approach to achieving that goal (Berkeley, Bender, Gregg Peaster, & Saunders, 2009). In MTSS models, all children receive Tier 1 classroom instruction focused on important academic skills. Children are screened periodically to identify those who are falling behind developmentally. Instruction, often categorized into Tier 2 and Tier 3 levels of support, is provided as appropriate

[a]University of South Florida, Tampa
[b]University of Missouri, Columbia
[c]University of Kansas, Kansas City
[d]The Ohio State University, Columbus
[e]Northern Arizona University, Flagstaff

Correspondence to Howard Goldstein: hgoldstein@usf.edu

Editor: Rhea Paul
Associate Editor: Shelley Gray

Received June 26, 2015
Revision received September 4, 2015
Accepted September 10, 2015
DOI: 10.1044/2015_JSLHR-L-15-0227

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

to levels of need. Progress is monitored so that adjustments can be made in the tiers of support children receive. There is evidence that MTSS can be effective in improving child outcomes (Burns, Appleton, & Stehouwer, 2005), and a few studies have examined MTSS in early childhood settings (Gettinger & Stoiber, 2008; VanDerHeyden, Snyder, Broussard, & Ramsdell, 2008; VanDerHeyden, Witt, & Gilbertson, 2007). To effectively implement MTSS to address language deficits, there is a need for efficacious Tier 2 and Tier 3 interventions. Tier 2 interventions typically supplement Tier 1 instruction and often are delivered by classroom staff in small groups. Because many children in a classroom may require Tier 2 language intervention (Greenwood et al., 2013), it is especially important that interventions can be delivered with high fidelity using existing classroom resources.

Tiered vocabulary programs have been implemented with preschoolers and kindergartners (Loftus, Coyne, McCoach, & Zipoli, 2010; Pullen, Tuckwiller, Konold, Maynard, & Coyne, 2010; Zucker, Solari, Landry, & Swank, 2013). These studies found that at-risk students made greater vocabulary gains when a second tier of instruction was added. This may not be surprising, as observational studies have found limited instructional support in early childhood settings (Greenwood et al., 2013). MTSS research in the primary grades can inform the design of Tier 2 interventions to maximize learning (Gersten et al., 2008). Effective instruction needs to be explicit, teach important instructional targets, and support learning through scaffolding and many opportunities to respond (Foorman & Torgesen, 2001; Snow, Burns, & Griffin, 1998). Vocabulary interventions that incorporate some of these components have produced moderate to large effects on proximal measures of student learning (Beck & McKeown, 2007; Coyne, McCoach, & Kapp, 2007; Johnson & Yeates, 2006; Loftus et al., 2010; Marulis & Neuman, 2010; Mol, Bus, & de Jong, 2009; Neuman, Newman, & Dwyer, 2011).

Tier 2 instruction needs to be characterized by high fidelity and high feasibility to sustain implementation in authentic early childhood education settings. Fidelity of implementation—the extent to which a treatment is implemented as intended—is a critical consideration for effective intervention. Researchers have found that higher levels of fidelity produce stronger treatment effects (Pas & Bradshaw, 2012; Zucker et al., 2013). When intervention is implemented by educational staff, fidelity of implementation is often lower than when intervention is implemented by researchers (Hulleman & Cordray, 2009; Mol et al., 2009). Dickinson (2011) reported fidelity of just 38% of instructional elements during a shared storybook reading intervention following teacher training. Other research groups have reported similarly discouraging findings (Justice, Mashburn, Hamre, & Pianta, 2008). Within a particular treatment, particular components may be more or less easy to implement in classrooms. Adherence to procedures of an intervention (e.g., providing specific activities, using prescribed materials) is frequently high, whereas the use of instructional techniques and processes (e.g., language modeling) is much lower (Justice et al., 2008; Pence, Justice, & Wiggins, 2008). Zucker et al. (2013) found that implementation fidelity was higher for the first tier of instruction than for the second tier. The authors hypothesized that the shared storybook reading context and strategies utilized in Tier 1 were familiar to teachers and thus easy to implement with fidelity. The explicit, extended vocabulary teaching expected in Tier 2 was less familiar and more challenging. Consideration of implementation feasibility prompted the development and evaluation of Story Friends, a Tier 2 pre-K curriculum that delivers key instructional components using automated prerecorded lessons.

Goldstein and colleagues conducted a series of studies to develop, refine, and examine the effects of the Story Friends curriculum, a Tier 2 intervention in which vocabulary and comprehension instruction is embedded in specially designed storybooks and prerecorded (Greenwood et al., in press; Kelley, Goldstein, Spencer, & Sherman, 2015; Spencer, Goldstein, Sherman, et al., 2012). A brief description of the program is provided below; additional discussion of the curriculum design and details of the iterative development process are available in Kelley and Goldstein (2014).

Story Friends is a supplemental preschool program designed with feasibility of high implementation fidelity. Rather than a Tier 2 program specific to a single early childhood curriculum, the program is designed to accompany a variety of Tier 1 curricula. The program targets two important skills within the domain of oral language: vocabulary knowledge and language comprehension. To address vocabulary knowledge, sophisticated vocabulary words were targeted using explicit instruction. Vocabulary targets were selected by applying Beck and McKeown's (2007; Beck, McKeown, & Kucan, 2002) recommendations to teach challenging, high-utility vocabulary words that occur frequently in the language of adult language users and in written text. Beck and colleagues argued that these sophisticated vocabulary words, rather than commonly used words likely to be acquired without instruction, are the most appropriate targets for explicit instruction. For the Story Friends program, we applied the following criteria to select words: (a) words that were unlikely to be familiar to preschool children with limited vocabulary; (b) words that were likely to occur frequently in sophisticated spoken and written language; (c) words that could be defined with a simple, child-friendly explanation; and (d) words that could be supported in the story context with explicit instruction (Spencer, Goldstein, & Kaminski, 2012). Systematic, explicit instruction for each vocabulary target was designed to include a reference to the story context; a simple, child-friendly definition; and connections between the word's meaning and children's everyday experiences. A recorded narrator provided several opportunities for children to say the word and the definition and then modeled correct responses. Many of these components of explicit instruction have been identified in other vocabulary intervention studies (e.g., Beck & McKeown, 2007; Coyne et al., 2007).

To address language comprehension, Story Friends targets answering questions about stories. Our preliminary studies of Story Friends found that most preschoolers were able to answer literal questions. Therefore, inferential questions for which the answer was not explicitly stated in the story text were the focus of instruction. Inferential language has been shown to contribute to reading comprehension (Cain, Oakhill, & Bryant, 2004; Cain, Oakhill, & Lemmon, 2004). Inferential questions included poststory predictions (e.g., "Do you think the Jungle Friends will go to the beach again? Why/why not?"), questions about character emotions (e.g., "Why is Ellie happy?"), and questions about character actions (e.g., "Why did Leo help Marquez?"). Embedded lessons for story questions included a model of an appropriate response and a "think aloud" explanation of the response. This approach to teaching inferential language in the context of storybooks was informed by the work of van Kleeck and colleagues (van Kleeck, 2006, 2008; van Kleeck, van der Woude, & Hammett, 2006). Appendix 1 includes sample scripts for vocabulary and comprehension lessons.

To ensure high-fidelity implementation in classroom settings, Story Friends makes use of an innovative automated approach. Stories and embedded lessons are prerecorded; children listen through headphones in small groups and follow along in accompanying storybooks. The automated approach ensures that key elements of the carefully designed, explicit instruction (e.g., exposure to instructional targets, systematic instructional language, multiple opportunities to respond) are delivered consistently. Rather than requiring extensive professional development or dedicated resources, Story Friends requires minimal training of educational staff and can be incorporated into a common context in early childhood classrooms (e.g., the listening center). Procedural fidelity (e.g., setting up materials) was the responsibility of educational staff, whereas fidelity of instructional components was accomplished via the automated instruction.

The efficacy of the automated approach incorporated in Story Friends has been examined in a series of studies (Greenwood et al., in press; Kelley et al., 2015; Spencer, Goldstein, Sherman, et al., 2012). Across studies, large effect sizes were demonstrated on proximal measures of vocabulary. During the iterative development of Story Friends, the curriculum was modified on the basis of an analysis of learning patterns. Children subsequently were able to define more of the targeted words. No significant differences were found on the Peabody Picture Vocabulary Test–Fourth Edition (PPVT-IV; Dunn & Dunn, 2007), however, as the words targeted are not ones assessed in standardized tests except for words that show up at higher age levels.

Proximal measures that tested responses to story questions targeted in the program showed clear effects, although some correct responding at pretesting after hearing the story once left limited room for improvement. To assess whether children could demonstrate generalized ability to answer questions about stories, the Assessment of Story Comprehension (ASC; Spencer, Goldstein, Kelley, Sherman, & McCune, 2015) was developed. Participants demonstrated improvement in their ability to answer inferential questions relative to a comparison group (Kelley et al., 2015).

In these prior studies, learning in the Story Friends condition was compared to that in business-as-usual conditions. To determine whether the embedded lessons were responsible for these robust effects, it was necessary to test the effects of hearing the words in the stories without embedded lessons. Thus, Kelley et al. (2015) included untaught words in their single-case experimental designs and tested children on these words contained in the stories that the explicit instruction did not target. Learning of untaught words through this exposure alone rarely resulted in learning. To confirm this finding in the current randomized control trial, a storybook condition that provides exposure to words and stories in the same small-group format offers a more appropriate comparison condition than a business-as-usual control. This has the potential of elucidating the role of our hypothesized active ingredient (explicit, embedded instruction) in the intervention. Another benefit of an exposure-only comparison is that it makes the vocabulary outcome measures "fair" to the comparison group; children in the comparison condition are exposed to the words on the measures. This would not be the case with a business-as-usual control group, as children would not have opportunities to learn the tested words in their Tier 1 experiences.

In previous evaluations of Story Friends (Greenwood et al., in press; Kelley et al., 2015), research staff was responsible for facilitating the listening center. Fidelity of implementation was high, even when less-experienced undergraduate students served as facilitators. Pilot studies of educator-led Story Friends also indicated that the program had feasibility of high-fidelity implementation; paraprofessionals were able to manage the listening center procedures with minimal training. In the current study, educational staff in the classroom implemented Story Friends, a critical next step in examining the feasibility of the program.

The purpose of this study was to determine whether embedded lessons (experimental group) produced more learning of vocabulary and comprehension skills than listening to the same stories without embedded lessons (comparison group) when implemented in the context of small-group listening centers in preschool classrooms by educational staff. We hypothesized that the Story Friends curriculum would produce significantly better outcomes on measures of vocabulary learning and question answering than simply listening to the stories. Given previous research on the influence of fidelity and preintervention language skills on treatment effects (e.g., Penno, Wilkinson, & Moore, 2002; Zucker et al., 2013), moderator effects of these variables were examined. Higher doses and preintervention language skills were expected to predict improved learning. The following research questions were addressed:

1. Controlling for children's language pretest scores, what are the effects of the Story Friends program on proximal measures of vocabulary and comprehension learning?
2. What are the effects of the Story Friends program on distal language measures?

3. Are observed treatment effects in the experimental group moderated by the number of instructional book listens attained or by the children's pretest skills on the Clinical Evaluation of Language Fundamentals (CELF-P; Wiig, Secord, & Semel, 2004)?

Method

Experimental Design and Power Estimate

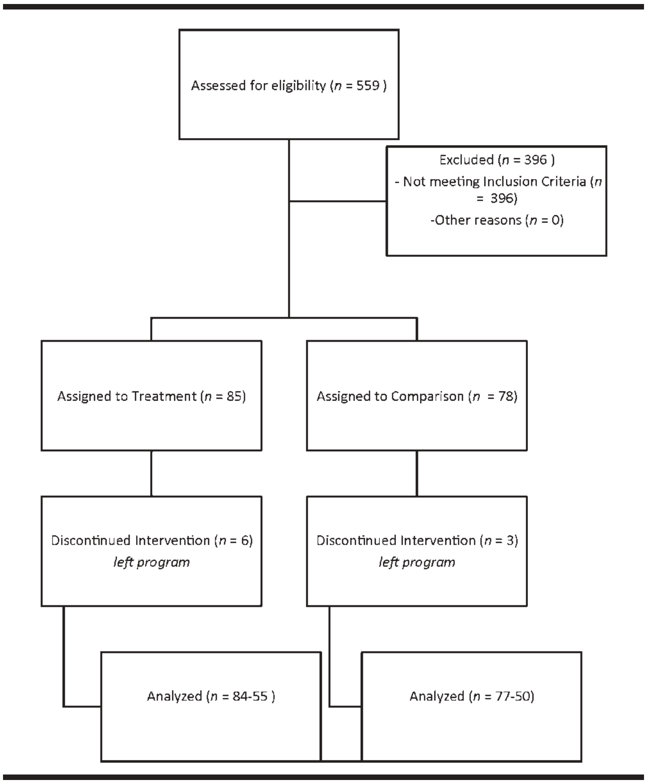
A cluster randomized design with children nested in classrooms was used to compare the effects of Story Friends with and without embedded lessons on measures of vocabulary learning and comprehension (Rhoads, 2011). Power estimates were computed to determine the number of classrooms needed to detect treatment effects with power of .80. The alpha level was set at .05 and intracluster correlations (ICCs) were set at .20 on the basis of a review of education intervention studies (Hedges & Hedberg, 2007). A child-level covariate was included in the power analyses to capture probable correlations between child-level variables (e.g., pretreatment performance on language measures such as the CELF-P) and treatment effects. The child-level covariate was set at $R^2 = .52$ on the basis of previous research with a similar sample (Greenwood et al., in press). The power estimates indicated that a sufficiently powered study would include 30 classrooms with four to six children per classroom. Attrition at the child level was accounted for with the addition of 15% more children, but attrition was expected to be unlikely at the classroom level.

Participants

The study was conducted in public school pre-K classrooms in two urban sites: Columbus, Ohio and Kansas City, Kansas. There were 24 classrooms in Ohio and eight classrooms in Kansas. Classrooms served primarily students from families with low incomes. The Ohio programs were full day 5 days per week, and the Kansas programs were half day 4 days per week. At each site, classrooms were randomly assigned to the experimental and comparison conditions. The flow of participants through the experiment is shown in Figure 1.

To identify participants at increased risk and likely to benefit from Tier 2 intervention, we considered information from the Individual Growth and Development Indicators (IGDI) Oral Language screening measures (Picture Naming and Which One Doesn't Belong; Bradfield et al., 2014) and a standardized norm-referenced measure (PPVT-IV). The goal was to identify six participants per classroom. All eligible children completed the two IGDIs. On the Picture Naming IGDI, children were asked to verbally label a set of 15 pictures. On the Which One Doesn't Belong IGDI, children were presented with a set of three pictures (e.g., shirt, pants, and boat) and asked to point to the picture that did not belong. A cutoff score provided by the IGDI developers was applied; all children with scores above the cutoff score on both IGDIs were excluded as Tier 2 candidates. Children with scores below the cutoff score on one or both IGDIs were given the PPVT-IV. The first six children with standard scores between 1.5 and 1.0 *SD*s below the mean (standard score = 78–85) were included. In classrooms in which fewer than six children had scores between 78 and 85, we extended the range up and down to include sufficient participants (standard score range = 71–96).

Figure 1. CONSORT diagram showing the flow of participants throughout the experiment.


Assessed for eligibility (*n* = 559)
Excluded (*n* = 396)
- Not meeting Inclusion Criteria (*n* = 396)
-Other reasons (*n* = 0)
Assigned to Treatment (*n* = 85)
Assigned to Comparison (*n* = 78)
Discontinued Intervention (*n* = 6)
left program
Discontinued Intervention (*n* = 3)
left program
Analyzed (*n* = 84-55)
Analyzed (*n* = 77-50)

A total of 163 students were enrolled in the study, with an average of five students per classroom. As can be seen in Table 1, statistical tests revealed no significant differences between the groups on demographic, developmental, or attrition variables. Children averaged 58 months of age at pretest. Standard scores averaged 83.9 and 84.2 at pretest for the PPVT-IV and 83.1 and 80.3 for the CELF-P for the experimental and comparison groups, respectively. The percentages of children with individualized education programs averaged 2.5% and 5.1% of the samples, and children who were English language learners averaged 18.8% and 17.9% of the samples for the experimental and comparison groups, respectively. Child attrition totaled 6% over the course of planned data collection.

Setting and Procedures

In all classrooms, children participated in small-group (approximately three children) listening centers. The grouping of children was flexible and determined by the teacher. Classroom staff, a teacher, or a teacher's aide was responsible for facilitating the listening center (e.g., helping children keep

Table 1. Sample characteristics by experimental groups and sites.

		Group		Statistic		
Variable	Site	Experimental	Comparison	Estimate	*df*	*p*
Children (*n*)	Ohio	60	54			
	Kansas	25	24			
	All	85	78			
Classrooms (*n*)	Ohio	12	12			
	Kansas	4	4			
	All	16	16			
Age at start (months), *M* (*SD*)	Ohio	57.4 (3.38)	57.8 (3.44)	*t* = 0.56	112	.575
	Kansas	57.5 (3.03)	58.9 (3.78)	*t* = 1.44	47	.156
	All	57.4 (3.26)	58.1 (3.56)	*t* = 1.28	161	.203
Individualized education program status (% yes)	Ohio	1.7	3.70	χ^2 = 0.46	1	.498
	Kansas	5.3	8.30	χ^2 = 0.15	1	.695
	All	2.5	5.10	χ^2 = 0.72	1	.396
English language learner status (% non-English)	Kansas	48.0	41.70	χ^2 = 0.20	1	.656
	All	18.8	17.90	χ^2 = 0.02	1	.886
CELF-P pretest, *M* (*SD*)	All	83.10 (11.07)	80.32 (12.56)	*t* = 1.48	153	.147
CELF-P posttest, *M* (*SD*)	All	88.94 (10.12)	87.64 (11.62)	*t* = 0.74	153	.461
PPVT-IV pretest, *M* (*SD*)	All	83.90 (5.32)	84.18 (5.70)	*t* = 0.32	153	.748
PPVT-IV posttest, *M* (*SD*)	All	93.29 (8.31)	91.83 (9.22)	*t* = 1.04	153	.301
Sample size by unit, *M* (*SD*)						
1		84	77			
2		79	76			
3		79	76			
4		79	76			
5		79	74			
6		55	50			
Omnibus				χ^2 = 0.12	5	.999

Note. Standard deviation given in parentheses. CELF-P = Clinical Evaluation of Language Fundamentals–Preschool; PPVT-IV = Peabody Picture Vocabulary Test–Fourth Edition.

headphones on or turn pages). Children followed along in storybooks, listened to accompanying prerecorded audio, and responded to questions and prompts to say words and definitions. Research staff supported teacher implementation by providing materials, assisting with scheduling, and troubleshooting equipment and monitored fidelity. Research staff was responsible for the administration and scoring of child assessments.

The Story Friends curriculum included two book series (*Forest Friends* and *Jungle Friends*). Each book series had an introductory book designed to familiarize participants with characters and procedures, nine instructional books, and three review books. The nine instructional books in each series were divided into three units of three instructional books plus a review book. For the introductory and instructional books, children listened to each book three times in a week. Due to constraints of scheduling assessments, repeated readings of the review book were not possible. On those weeks, children received one session. Appendix 2 provides an overview of the timeline for intervention and assessment procedures.

In the experimental condition, the storybooks and prerecorded audio included embedded lessons on challenging vocabulary words and story questions (cf. Spencer, Goldstein, Sherman, et al., 2012). Each instructional book included two embedded lessons for each of two challenging vocabulary words (e.g., *enormous, brave*) and one embedded lesson for each of three inferential story questions (e.g., "Why was Leo sad?"). The review book included brief lessons for each of the six vocabulary words included in the three instructional books in that unit. The prerecorded storybooks were 9 to 12 min in length. Because children heard storybooks three times per week, they received 30 to 40 min of additional instruction per week. Introductory books did not include instruction, and children listened to review books just once. Thus, across the six units, participants received about 10 hr total of instruction. In Kansas, participants completed five of the six units due to school year scheduling; the sixth unit was completed in Ohio only. Across the three listens to the instructional books and the single listen to the review books, children received seven embedded lessons for each challenging vocabulary word and three lessons for each story question.

In the comparison condition, participants were exposed to the same stories containing the same vocabulary words, but with no embedded lessons. Participants followed the same procedures (e.g., repeated readings of the instructional books), but the length of the storybooks was slightly shorter without the embedded lessons.

Outcome Measures

Researcher-developed and standardized measures were administered to identify participants, measure proximal and distal outcomes, monitor fidelity of implementation, and assess teacher satisfaction. Learning of instructional targets was assessed using two researcher-developed measures:

the Unit Vocabulary Test (UVT) and the ASC. Trained research staff administered the UVT and the ASC individually to participants.

The UVT was administered prior to and immediately following each unit of instruction (i.e., three instructional books and one review book) and tested each of the six vocabulary targets for the unit. A definitional task was selected for the UVT to provide a rigorous measure of vocabulary learning. Multiple-choice or picture-pointing tasks could be correct by chance, whereas a definitional task had no chance component. The definitional task also allowed for incremental scoring to be sensitive to partial word knowledge.

For each item on the UVT, participants were asked an open-ended question: "What does [target word] mean?" If participants did not provide a correct definition, a standard cloze prompt ("[target word] means …") was delivered one time. Participant responses were transcribed in real time; audio recordings were collected for reliability purposes. Trained research assistants scored the UVT using detailed scoring guides. Each response could receive 2, 1, or 0 points. A 2-point response was the taught definition, a synonym, or another complete response. A 1-point response was a response that indicated some knowledge of the word, such as an example from the story (e.g., "An elephant is enormous"). A 0-point response was an unrelated response, an "I don't know" response, or no response. Scoring guides included multiple sample answers for each scoring category for each item.

The ASC served as an outcome measure for the comprehension domain. The ASC is a curriculum-based measure designed to assess children's ability to answer questions about stories. The ASC has evidence of good psychometric properties, including high internal consistency (Cronbach's $\alpha = .96$), test–retest reliability ($r = .82$), and strong correlations with measures of oral language (e.g., $r = .81$ with CELF-P), making it an appropriate tool for this investigation.

Trained research staff administered the ASC individually, and participants completed one ASC story at each of seven time points (preintervention and after each unit). The ASC includes nine parallel forms; a standard sequence of forms was used for all participants. To administer the ASC, the examiner reads a short story and asks the child to respond to a series of questions about the story. Questions include three literal questions about key story events (e.g., "What was Danny doing in this story?"); a question about a vocabulary word in the story (e.g., "What does *injured* mean?"); and four inferential questions about character actions (e.g., "Why did Danny's mom give him a hug?"), character emotions (e.g., "Why was Danny sad?"), or post-story predictions (e.g., "Do you think Danny will ride his bike down the big hill again?"). The inferential question types were similar to those targeted in Story Friends. The ASC is scored on a scale of 1 to 17. Literal and inferential questions were scored on a 3-point scale. A score of 2 was awarded for a complete and clear response (e.g., "He will get hurt"), a score of 1 was awarded for a correct response that was incomplete or unclear (e.g., "Fall down"), and a score of 0 was awarded for an incorrect or unrelated response (e.g., "Watch TV"). The vocabulary item was scored on a scale of 0 to 3.

The UVT and ASC served as proximal measures of student learning; two standardized, normative language measures—the PPVT-IV and the CELF-P—were administered at pretest and posttest to measure distal outcomes. Information from the PPVT-IV was also considered for participant identification. Participants completed three subscales of the CELF-P (Sentence Structure, Word Structure, and Expressive Vocabulary) to obtain a Core Language score, which has a normative mean of 100 and an *SD* of 15.

Dosage

The attendance of children in both groups at all scheduled sessions was monitored as an indicator of exposure to each curriculum condition. The intended dosage was 10 listens per unit (three listens each to three instructional books and one listen to the review book). Reasons for missing these opportunities to learn included absences from school for sickness or other excused reasons and cancellations due to school schedule changes and winter weather. Make-up sessions were arranged whenever possible. On rare occasions, a child missed an entire unit due to an extended absence. Overall, dosage was high in both groups and at both sites. In Ohio, the experimental group received an average of 8.79 listens per unit (*SD* = 1.56, range = 0–10), and the comparison group participated in an average of 8.49 listens per unit (*SD* = 1.90, range = 1–10). Means were similar in Kansas: 8.83 listens per unit (*SD* = 1.67, range = 0–10) for the experimental group and 8.83 listens per unit (*SD* = 1.43, range = 5–10) for the comparison group. Across sites, the number of listens declined slightly in the experimental group at a rate of –0.13 listens per unit. There was no decline in the number of listens in the comparison group.

Implementation Fidelity

Research staff assessed fidelity of intervention using checklists that included six items reflecting critical components of the intervention (i.e., each child had headphones, each child had a book, facilitator had headphones, correct audio was playing and functioning properly, entire audio was played, listening center was quiet with few distractions). Observations were conducted every 1 to 2 weeks. Research staff observed each listening center (e.g., two listening center observations per classroom per visit). The average number of observations per classrooms was 18 for treatment classrooms and 10 for comparison classrooms.

Rigorous standards were set for the fidelity of implementation of the automated interventions (at least 90%). Overall fidelity of implementation was quite high: 94.9% (range = 33%–100%) across all classrooms. In treatment classrooms, fidelity of implementation was 93.9% (33%–100%). In comparison classrooms, fidelity of implementation was 96.6% (67%–100%). Instances of low fidelity were rare; most often these were isolated incidents in which the listening center was interrupted (e.g., fire drill, other change in schedule). Across all observations, the item on the fidelity checklist that was most often missing was "listening center was quiet with few distractions"; only 84% of observations indicated that this was the case.

Research staff completed training and "check-out" procedures for all measures prior to administration. Audio recordings of assessment sessions were used to conduct fidelity checks for administration of the measures throughout the study (30% of measures at each unit of assessment). To examine fidelity of test administration, trained research assistants listened to audio recordings and completed procedural checklists specific to each measure (e.g., assessor delivered items exactly as written, provided standard prompts). The criterion for fidelity of assessment was 85% of items on the checklist. All observed sessions (374 observations across sites) exceeded this criterion; overall fidelity was very high. For the UVT, average fidelity of administration was 99.97% (range = 93%–100%). For the ASC, average fidelity of administration was 100% (range = 85%–100%).

Scoring Reliability

A primary scorer, a trained member of the research team, scored all measures. At each site, a second research team member scored approximately one third of all the assessments for the purpose of evaluating scoring reliability. Detailed scoring guides were created to ensure reliable scoring. For both the UVT and the ASC, scoring rubrics and sample responses were created for each item. For the UVT, scorers were blinded to condition (treatment, comparison) and to the unit of assessment (pre–post). For the ASC, scorers were blinded to condition. Because the ASC forms were administered in a standard order, it was not possible for scorers to be blind to unit of assessment.

On both measures, an item-by-item comparison was made to determine agreement or disagreement. Scoring reliability was calculated by dividing the total number of agreements by the total number of agreements plus disagreements and multiplying by 100. Kappa coefficients also were calculated.

For the UVT, the high number of 0-point answers (e.g., "I don't know" or no response) at pretest made scoring agreement more likely. Thus, we examined agreement separately for pretest and posttest. Mean agreement was 98% at pretest ($\kappa = .84$) and 97% at posttest ($\kappa = .91$). On the ASC, mean agreement was 93% ($\kappa = .88$) across all units of assessment. Cross-site scoring reliability was examined for 10% of all assessments samples across units of assessment. Mean agreement was 99% ($\kappa = .99$) on the UVT and 93% ($\kappa = .88$) on the ASC.

Social Validity Assessment

Following intervention, teachers and teacher's aides were asked to complete anonymous consumer satisfaction surveys (approximately 93% return rate). For the purposes of evaluating Story Friends, only the survey results from the teachers ($n = 20$) and teachers' aides ($n = 12$) in the treatment condition who completed the survey were examined. The survey consisted of 17 items related to satisfaction with the intervention (e.g., "The listening centers are worth the time and effort and I would implement them again sometime in the future"), ease of implementation (e.g., "I could fit the listening center into my classroom schedule three times each week without much difficulty"), and participation in the research study (e.g., "I had the support I needed from the research team to conduct the listening center"). Teachers responded on a scale of 1 (*strongly disagree*) to 6 (*strongly agree*). The survey included several open-ended questions (e.g., "If you could change one thing about the intervention, what would you change?").

Results

Data Analysis

Mixed regression modeling was used to answer research question 1 regarding the observed effects on vocabulary and comprehension learning. Estimates of the between-children and between-classrooms variation are included in the models to account for the clustering of multiple observations across time within each child and across multiple children within each classroom (Snijders & Bosker, 2012). To control for children's pre-existing language differences at start, pretest scores on the CELF-P and each instructional unit were included in the model predicting UVT and ASC scores. Next, differences by treatment group (experimental vs. comparison) were tested in the model first for the main effect of entire series (grand mean collapsed over units) and second for the effects across the six instructional units, forming a multilevel growth model. Growth over time was indicated by the interaction of group (two) by unit (six).

In these models, estimates for first-unit score (i.e., intercept) and linear growth across units were allowed to vary across children. To account for the nesting of children within classrooms, intercept and growth estimates were allowed to vary across classrooms such that modeling accounted for classroom-level differences in first-unit scores and growth not explained by control and focal predictors. Because these models did not assume a particular data structure for the number of data points and because prior data for the Kansas children were included, the analyses were not influenced by the missing Kansas data at unit 6.

In addition to model estimates and p values for each predictor, effect size estimates were calculated. Cohen's f^2 (Cohen, 1988) was used as an appropriate local effect size for representing each variable's impact in the multivariate mixed-effects regression models; Cohen's f^2 values of .02, .15, and .35 reflect small, medium, and large effects, respectively. Derivation of the marginal semipartial R^2 values necessary for computation of Cohen's f^2 was calculated after Nakagawa and Schielzeth (2013).

Because the raw data distribution of UVT for vocabulary was positively skewed due to a large numbers of zeroes and a long positive tail, Poisson regression was used to fit the distribution (Agresti, 2007). Poisson regressions with log links between the predictors and the dependent variable were used at the core of the conducted multilevel growth modeling. The distribution of ASC comprehension scores was generally normal, with no substantial skewness (0.318) or kurtosis (–0.555) present; therefore, general linear mixed modeling was appropriate for modeling comprehension.

To answer research question 2 regarding the effects on posttest gains in PPVT-IV and CELF-P scores, linear mixed models were used. Because each child had one posttest score for each measure, only nesting of children within classrooms was accounted for by the inclusion of a random intercept. Inclusion of PPVT-IV and CELF-P pretest scores as predictors was used to sort out associated variance, allowing group to uniquely predict gains from pretest to posttest on the outcomes. Research question 3 investigated moderation of the growth within the experiment depending on the child's starting language skill (CELF-P) and dosage of exposure to the intervention (total number of listens). The CELF-P × Unit × Number of Instructional Book Listens interaction was examined as a predictor of UVT and ASC scores.

Effects on Vocabulary Learning

Vocabulary learning is reported in word points per unit, with a maximum of 12 reflecting complete knowledge of all six novel words per unit. As can be seen in Table 2, controlling for children's initial differences in pretest CELF-P language and unit pretest vocabulary knowledge, the main effect of group on children's vocabulary learning was large and statistically significant (group $\beta = 1.58$, $p < .001$, Cohen's $f^2 = .70$). The experimental group grew from a pretest mean of 0.60 word point ($SD = 0.25$) to a posttest mean of 4.00 ($SD = 1.45$), a mean gain of 3.40 word points per unit. In contrast, the comparison group grew from a pretest mean of 0.50 word point ($SD = 0.18$) to a posttest mean of only 0.80 word point ($SD = 0.36$), a mean gain of only a fraction of a word point (0.32). Learning in the experimental group was always significantly greater than that in the comparison group, ranging from 7.2 times greater at the first unit ($\beta = 2.106$, $\exp(\beta) = 8.218$, $SE = 0.146$, $p < .001$) to 1.9 times greater at the last unit ($\beta = 1.048$, $\exp(\beta) = 2.853$, $SE = 0.184$, $p < .001$). Effect sizes were medium on the basis of Cohen's f^2 value of .17 for the variance explained by the interaction between group and unit and for the main effect of group alone ($f^2 = .23$). In contrast, the comparison group's gains were negligible and nonsignificant (i.e., a rate of 0% per unit)—$\beta = 0.016$, $\exp(\beta) = 1.016$, $SE = 0.036$, $p = .67$.

In Figure 2, low pretest scores for each unit are evident for both the experimental and comparison groups. Although equally low post-test scores are evident for the comparison group, the experimental group demonstrated clear improvements in word learning for each unit, averaging 3.4 word points. Because new words were introduced at each unit, means for each unit were not expected to grow. However, the interaction of experimental group by curriculum unit over time was statistically significant (Group × Unit $\beta = -0.210$, $SE = 0.042$, $p < .001$), reflecting an unexpected declining trend over time. ICCs in these analyses for children were ICC = .103 and for classroom were ICC = .066, indicating that differences between children explained 10.3% of the variance in vocabulary posttest scores, whereas differences between classrooms explained 6.6% of the variance in vocabulary posttest scores.

Effects on Comprehension Learning

After controlling for initial differences, there were no significant findings for the effects of groups or the Group × Unit interaction on the ASC total score. Children in the experimental group achieved a mean of 8.3 ($SD = 4.0$) at the last assessment versus 7.2 ($SD = 3.9$) in the comparison

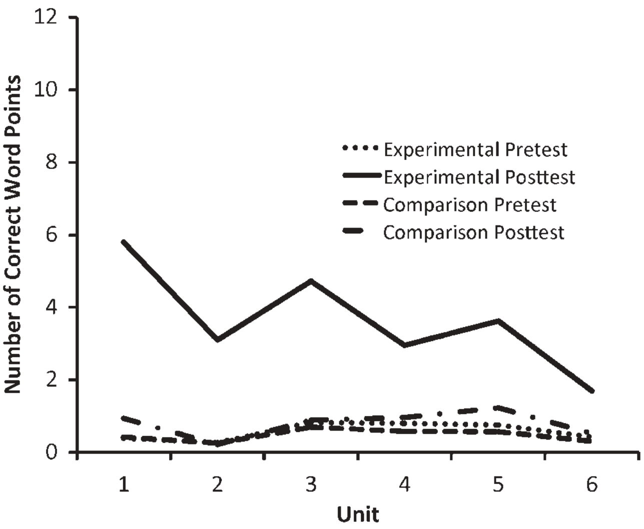
Table 2. Mixed Poisson regression results for vocabulary learning and mixed linear regression results for comprehension

	Vocabulary learning				Comprehension		
Model/fixed effects	$\hat{\beta}$ (*SE*)	*p*	**Effect size**	**Increase**[a] **(%)**	$\hat{\beta}$ (*SE*)	*p*	**Effect size**
Averaged across units							
Intercept	–2.702 (0.422)	<.001			–4.655 (1.205)	<.001	
Pretest score	0.105 (0.021)	<.001	.022	11.09			
CELF-P scale score	0.027 (0.005)	<.001	.116	2.69	0.130 (0.014)	<.001	0.206
Group[b]	1.580 (0.116)	<.001	.698	385.46	–0.230 (0.337)	.495	0.001
Growth across units							
Intercept	–2.822 (0.415)	<.001			–6.274 (0.147)	<.001	
Pretest score	0.136 (0.022)	<.001	.020	14.60			
CELF-P scale score	0.027 (0.005)	<.001	.130	2.75	0.147 (0.013)	<.001	0.316
Group[b]	2.023 (0.135)	<.001	.800	655.90	–0.071 (0.329)	.829	0.001
Unit[c]	0.013 (0.036)	.731	.046	1.26	0.109 (0.075)	.149	0.005
Group[b] × unit[c]	–0.210 (0.042)	<.001	.168	–18.94	–0.070 (0.108)	.514	0.002

Note. Effect size is given as Cohen's f^2. CELF-P = Clinical Evaluation of Language Fundamentals–Preschool.

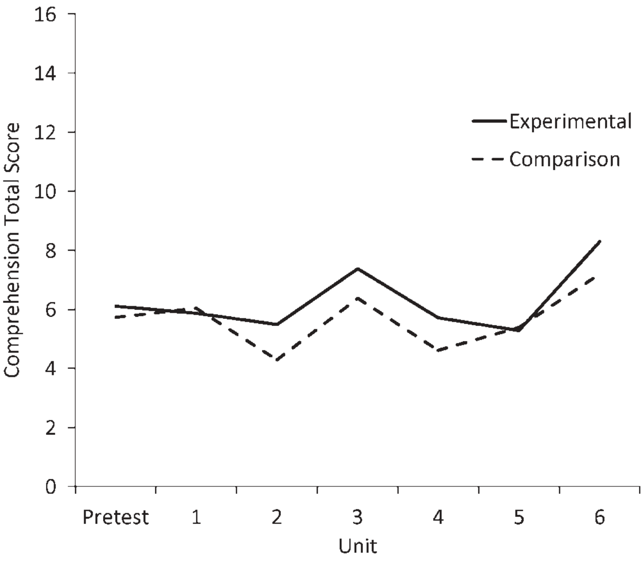
[a] Represents the percentage increase in word points learned per unit increase in the independent variable. [b] 0 = comparison, 1 = treatment. [c] Centered at unit 1.

Figure 2. Pretest and posttest vocabulary performance for experimental and comparison groups.



group—improvements of 2.2 and 1.5 points, respectively. Likewise, no significant differences between groups were seen at each unit (see Figure 3). However, a main effect of site (β = 1.818, *SE* = 0.428, *p* < .001) was noted, indicating that children in Ohio averaged 1.8 higher on the ASC Total Score compared with children in Kansas. ICCs were calculated for children and for classroom levels (child ICC = .214, class ICC = .207). Differences between children explained 21.4% of the variance in comprehension posttest scores, whereas differences between classrooms explained 20.7% of the variance in comprehension posttest scores.

Figure 3. Performance on the Assessment of Story Comprehension at pretest and after each unit for experimental and comparison groups.



Effects on Distal Language Measures

After accounting for initial differences in pretest scores, neither norm-referenced language measure (PPVT-IV, CELF-P) was influenced by group. However, both groups made gains in standard scores of approximately 0.5 *SD* (see Table 1).

Moderation in the Experimental Group

Because intervention effects on word points were significantly greater for the experimental group, moderation was explored. Analyses indicated that instructional book listens were not significant as a main effect. Their interaction with instructional unit and CELF-P was significant (p = .03) but represented a small effect size (Cohen's f^2 = .003). Children with higher CELF-P scores benefited most from additional book listens, but this positive effect of additional book listens in early units dropped over time after unit 2.

Social Validity Results

In general, teachers were satisfied with the intervention (*M* = 5.05 and *SD* = 1.15 on items related to satisfaction). Teachers agreed that children benefited from participating in the listening centers (*M* = 5.25, *SD* = 0.92) and that children enjoyed the listening center activities (*M* = 5.06, *SD* = 1.06). However, seven teachers disagreed with the statement that children enjoyed listening to the same book three times (*M* = 4.26, *SD* = 1.41). Teachers agreed that it was feasible to have an adult at the listening center (*M* = 5.16, *SD* = 0.90) and somewhat agreed that it was feasible to integrate the intervention into the daily classroom schedule and weekly routine (*M* = 4.84, *SD* = 1.21). Responses to the open-ended questions reflected similar sentiments. Many teachers commented that students benefited from the program and gave examples of vocabulary learning and gains in language skills (e.g., answering questions, talking more, using complete sentences). Teachers also commented on changes in student behavior related to paying attention and following directions. Teachers frequently expressed a desire to include all children in the classroom rather than just the identified participants. Although comments were generally quite positive, several teachers identified challenges to implementation, finding it difficult to keep children engaged (e.g., children wanted to participate in a different center) or to make time for the intervention (e.g., suggesting making the listening center sessions shorter or less frequent).

Discussion

The purpose of this study was to examine the effects of a Tier 2 curriculum implemented in preschool classrooms by educational staff on the vocabulary and comprehension learning of preschool children with limited oral language skills. We sought to determine whether stories with embedded lessons produced more learning of vocabulary and comprehension skills compared with the same stories without embedded lessons. We also examined intervention effects on distal measures of oral language skill and evaluated the contributions of dosage and preintervention language skills on treatment effects.

Results revealed that when storybooks included embedded lessons, vocabulary learning consistently exceeded the exposure comparison condition. The effect size associated with the differences between treatment and comparison groups was large. Participants in the treatment classrooms gained an average of 3.4 word points for each unit. Performance varied substantially among participants, with three children gaining only 1 word point and a high performer gaining 54 word points during the year. Participants in the comparison classrooms demonstrated almost no gains, averaging less than 2 word points over the course of the year. The virtual absence of learning of the same vocabulary words that were included in the story by the comparison group without the embedded lessons indicates that storybook listening alone has little effect on the learning of challenging words. This was the case with untaught words in our prior single-subject design experiments (Kelley & Goldstein, 2014; Kelley et al., 2015; Spencer, Goldstein, Sherman, et al., 2012) and was consistent with previous research (cf. Beck & McKeown, 2007; Justice, Meier, & Walpole, 2005); children appear to learn very little about new vocabulary words from simply hearing stories read aloud. The facilitative effects of storybook reading seem to accrue from instructional lessons, even when the interactions are with an audio-recorded narrator. Perhaps effects would be as strong or stronger if parents and teachers provided similar instruction and opportunities for children to respond. But that would require substantial training and preparation on their part and may sacrifice implementation fidelity.

One unexpected finding was the decline in vocabulary learning over the course of the academic year. The first-unit performance was surprisingly high and the final-unit performance was surprisingly low, which was largely responsible for the downward trend in vocabulary learning. The high vocabulary learning in the first unit may reflect a novelty effect. This was unexpected, however, as our prior research showed that some children seem to gain confidence in their ability to define challenging words over time. Indeed, a future line of research may seek to determine how many weeks of limited progress in small-group Tier 2 intervention provides an indication of the need for additional (Tier 3) support. Some variability among units may be due to competing activities that preoccupy children or teachers around holidays and the end of the school year. Although consistent word-selection criteria were applied across the stories, differences in word difficulty may have contributed to variability among units. Factors such as familiarity (Gray & Brinkley, 2011), word type (Alt & Plante, 2006), phonotactic probability, and neighborhood density (Storkel & Hoover, 2011) have been shown to contribute to differences in word learning success in experimental tasks; we did not control for or examine these variables. Likewise, systematic instructional language may contribute to differences in lesson effectiveness. For example, some lessons likely had examples that were more interesting or meaningful to children than others. Book-level differences could be examined by counterbalancing the order of books in subsequent studies.

O'Donnell (2008) has outlined treatment variables that could explain variability in responding. Some of these variables (e.g., duration and dosage) were measured and examined, but others are more difficult to capture (e.g., engagement and responsiveness). Waning enthusiasm, monitoring, or positive feedback by teachers or decreased interest or engagement by the children represent possible explanations. Baker, Kupersmidt, Voegler-Lee, Arnold, and Willoughby (2010) found that implementation of a new intervention was predicted by teachers' perceptions of the program and by the support provided by the center. If an end-of-the-year decrease in learning is seen regularly in early childhood settings, these variables may be important to evaluate in future studies.

The effects of comprehension intervention did not differ between the two groups; only a slight upward trajectory was demonstrated. This was the one variable that showed an effect of site, with children in Ohio averaging almost 2 points higher than children in Kansas on the ASC. The lack of significant effects of intervention on comprehension was unexpected because prior research had shown significant effects for the inferential questions on the ASC (Kelley et al., 2015). We suspect that in Story Friends there are too few teaching trials and opportunities to respond. Van Kleeck et al. (2006) reported strong effects on literal and inferential language from an intensive intervention delivered individually. It appears that the "think aloud" instructional approach did not produce robust effects in the automated format, perhaps because the prerecorded narration does not provide contingent feedback.

Consistent with the recommendations of the National Reading Panel (2000), multiple measures of outcomes in the study included both specific, researcher-created measures (e.g., UVT) and standardized, norm-referenced measures of general language abilities (PPVT-IV and CELF-P). Although treatment effects were evident on the researcher-created measures, differential group effects were not seen for the PPVT-IV or the CELF-P. It may not be surprising that these measures were not sensitive to the effects of Story Friends, as there is no overlap in the words taught and those tested on the PPVT-IV and CELF-P. Indeed, words targeted in Story Friends would likely show up as items for older students in the PPVT-IV. Two recent meta-analyses of vocabulary intervention studies have reported significant effects on standardized measures. Marulis and Neuman (2010) found that large effects were evident on researcher-created measures of specific vocabulary taught ($g = 1.21$), whereas more moderate effects were observed on standardized measures ($g = 0.71$). In a meta-analysis of the effects of vocabulary instruction on comprehension that included older children, Elleman, Lindo, Morphy, and Compton (2009) reported an effect size of .29 for standardized measures of vocabulary. A careful analysis of the words taught in intervention studies is warranted to help us understand the potential relation between targeted vocabulary and vocabulary assessed on various standardized measures.

It is possible that a relatively low-intensity intervention such as Story Friends (approximately 10 hr total across the school year) is not sufficient to produce changes in global language abilities. Many studies of targeted vocabulary intervention for young children have not included measures of generalized outcomes (Beck & McKeown, 2007; Coyne et al., 2007; Loftus et al., 2010; Silverman, 2007). Those that have examined treatment effects on distal outcome measures often have failed to find effects (Coyne et al., 2010; Leung, 2008). For example, Neuman et al. (2011) found no group differences on the Woodcock–Johnson Picture Vocabulary Subtest after a supplemental vocabulary program implemented across the school year. In some studies, intervention programs that have targeted more global language skills (e.g., dialogic reading) have produced effects on distal measures (Hargrave & Sénéchal, 2000; Lonigan & Whitehurst, 1998); in other studies, no significant effects were reported (Lonigan, Anthony, Bloomfield, Dyer, & Samwel, 1999). Efforts devoted to changing instructional practice (e.g., professional development, collaborative models) and environments (e.g., literacy access) have sometimes resulted in improvements on distal language measures (Hadley, Simmerman, Long, & Luna, 2000; Neuman, 1999; Wasik, Bond, & Hindman, 2006). It appears that improving language abilities of young children to the extent that effects can be observed on generalized language measures is quite challenging; it is likely that intensive, sustained instruction is necessary.

Although no significant group differences on distal measures were observed, both groups showed improvement during the school year. Both groups' scores averaged more than 1 *SD* below the mean at pretest. On the PPVT-IV, the experimental standard score group mean increased by about 9.39 points and the comparison increased by 7.65 points. On the CELF-P, the experimental standard score group mean increased by 5.84 points and the comparison group increased by 7.32 points. It is likely that gains on these generalized measures are an indication of the contributions of general, Tier 1 pre-K instruction to language development. Greenwood et al. (2013) reported similar gains as the result of Tier 1 instruction across a variety of preschool settings for children at the Tier 2 performance level (5.1 points on the PPVT-IV; 7.4 standard score points on the CELF-P).

Moderation analyses revealed that treatment effects were relatively unaffected by classroom and site differences. ASC scores were a little higher in Ohio classrooms. One possible explanation is that difference may be due to preintervention language abilities. There were more children in Kansas who were English language learners. The ASC may have been more taxing for them because it requires some expressive language ability to produce responses.

Pretest scores on normed language tests did not moderate vocabulary learning, nor was there a main effect for instructional book listens. This is probably related to the small range in the number of listens, as implementation was high throughout the study, with only a small decline as the study progressed. Although we found a Number of Listens × Instructional Units × CELF-P interaction, this was a small effect (Cohen's $f^2 = .003$). There was a positive effect of additional book listens for higher CELF-P scores that was evident only in the early units.

The lack of variation in instructional dosage represents a strength of this study. The teachers and their aides achieved exceptionally high implementation fidelity. Observations indicated that in the vast majority of sessions, all key elements of the intervention procedure were in place. Fidelity of 95% was achieved, which is substantially higher than levels reported by other research groups (e.g., 38% fidelity observed by Dickinson, 2011), and 81% of children received eight to nine of the intended 10 listens of books per instructional unit. One way in which such high fidelity was achieved was by asking educational staff to be responsible only for procedural fidelity (e.g., ensuring that participants listened to the books) rather than for complex instructional approaches (e.g., explicit vocabulary instruction). This allowed educational staff to devote resources, such as professional development and planning time, to other classroom activities, including Tier 1 instruction. Previous studies have found that procedural aspects of an intervention program can be readily implemented with fidelity (Justice et al., 2008; Pence et al., 2008), and our intervention was designed with this in mind.

Other studies have used similar instructional targets (e.g., challenging vocabulary), measures (e.g., definitional tasks), and populations (e.g., children at risk due to limited vocabulary knowledge), but few studies have included all these components (Loftus et al., 2010; Pullen et al., 2010). In comparison with the interventions examined in these studies, Story Friends appears to have strong effects on student learning. Participants in the current study learned an average of 10.2 words out of a possible 36, representing 28.33% of words taught. In Loftus et al. (2010), at-risk children in the Tier 2 treatment group averaged a one-word advantage in definitions over the control group; they defined 12.5% of the eight words targeted after 240 min of instruction. Pullen et al. (2010) reported an impressive posttest definitional knowledge of on average 3.67 of the eight words targeted in Tier 2 (45.9%) after 200 min of instruction. Researchers implemented the intervention in the Pullen et al. and Loftus et al. studies.

Studies of teacher-implemented vocabulary interventions typically have examined classroom-wide (Tier 1) curricula. For example, Beck and McKeown (2007) found that participants receiving the Text Talk intervention learned 16% to 20% of 22 words. Neuman et al. (2011) found improvements after implementing the World of Words program; pretest scores on a picture-pointing task ranged from 61% to 78%, indicating that children knew many of the targeted vocabulary words prior to intervention. Scores increased to 77% to 88% of words known versus 69% to 79% for the control group. Biemiller and Boote (2006) reported that children learned 10% of word meanings after repeated exposure from storybook reading. This percentage increased to 22% when words were explained to children and to 41% when instruction was intense. A common finding in these studies is that children learn only a small percentage of the novel words taught. There certainly is room for improvement as we devise additional strategies that may aid mastery and use measures that may be more sensitive to gains in the depth of knowledge acquired.

This was the first study to examine all six units of Story Friends across an entire school year. There was substantial variability across units, with average learning of 45% of words taught in the first unit versus 11% of words taught in the last unit. In previous studies of Story Friends, gains in word learning were larger (56% of words taught in Kelley et al., 2015; 45% of words taught in Spencer, Goldstein, Sherman, et al., 2012). It is possible that the population sampled in the current study was somewhat different than previous samples. Although means on the PPVT-IV were similar (Standard Score = 83–84), the sample mean on the CELF-P was slightly lower (Standard Score = 80 vs. 86 and 89 in Spencer et al. and Kelley et al., respectively).

Conclusions and Future Directions

The purpose of the current investigation was to examine the effects of prerecorded, embedded lessons on the learning of vocabulary and story comprehension abilities of preschool children. Our findings suggest that this innovative automated approach has potential as an approach to provide supplemental vocabulary instruction and can be implemented with high fidelity in preschool classrooms. The comparison group received exposure to storybooks and words in the same automated, small-group context to provide a rigorous test of the embedded lessons. An important next step would be to compare prerecorded lessons with storybooks and lessons delivered "live" by teachers. If stronger effects can be achieved by live delivery, we should consider ways to support this practice in preschool classrooms, continuing to emphasize the need for high-fidelity implementation with minimal requirements for educational staff.

There was substantial variability in word learning among participants in the experimental group, ranging from 2% to 75% of words being learned (averaging 28%). We used performance on standardized screening and norm-referenced measures to identify participants with limited oral language as candidates for Tier 2 intervention. However, it is clear that the program varied in its effectiveness for individual participants. Future research should examine ways to match children to tiers of intervention. In an MTSS approach, information about learning in response to instruction could be used in combination with scores on standardized measures.

Future research should continue to investigate ways to enhance vocabulary learning and retention. It is possible that we are underestimating the extent of learning by using the definition task, as there may be long-term benefits even if the child develops partial knowledge of new vocabulary words. Repeated experiences with a word in a variety of supportive language contexts are likely necessary for the development of deep, long-term word knowledge. The vocabulary learning that occurs in Story Friends may be one important component of the word learning process. We ultimately need to evaluate the long-term effects of vocabulary instruction on later reading fluency and comprehension.

Our project is unique because we attempted to teach challenging vocabulary words, measure definitional vocabulary knowledge, and work with at-risk preschool children in classroom settings using existing educational staff over the course of an entire school year. Our findings contribute to an existing body of research on explicit vocabulary instruction for preschool children and extend this work in important ways. Shared storybook reading has been a popular context for embedded vocabulary instruction (Beck & McKeown, 2007; Coyne et al., 2007; Justice et al., 2005); however, few studies to date have examined the use of automated methods of delivery. This important innovation provides a method for implementation in classroom settings without placing excessive demands (e.g., extensive professional development, additional personnel) on existing resources. Given the well-documented challenges of high-fidelity implementation in authentic educational settings, there is a clear need for such intervention approaches.

Acknowledgments

This work was supported by the Institute of Education Sciences, U.S. Department of Education through Cooperative Agreement R324C080011, which supported the Center for Response to Intervention in Early Childhood. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We appreciate the assistance of the teaching staff, administrators, and children in early childhood settings in Kansas and Ohio.

References

Agresti, A. (2007). *An introduction to categorical data analysis.* New York, NY: Wiley.

Alt, M., & Plante, E. (2006). Factors that influence lexical and semantic fast mapping of young children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 49,* 941–954.

Baker, C. N., Kupersmidt, J. B., Voegler-Lee, M. E., Arnold, D. H., & Willoughby, M. T. (2010). Predicting teacher participation in a classroom-based, integrated preventive intervention for preschoolers. *Early Childhood Research Quarterly, 25,* 270–283. doi:10.1016/j.ecresq.2009.09.005

Beck, I., & McKeown, M. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *The Elementary School Journal, 107,* 251–272. doi:10.1086/511706

Beck, I., McKeown, M., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction.* New York, NY: Guilford.

Berkeley, S., Bender, W. N., Gregg Peaster, L., & Saunders, L. (2009). Implementation of response to intervention: A snapshot of progress. *Journal of Learning Disabilities, 42,* 85–95. doi:10.1177/0022219408326214

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology, 98,* 44–62.

Bishop, D., & Adams, C. (1990). A prospective study of the relationship between specific language impairment, phonological disorders, and reading achievement. *Journal of Child Psychology and Psychiatry, 31,* 1027–1050.

Bradfield, T. A., Besner, A. C., Wackerle-Hollman, A. K., Albano, A. D., Rodriguez, M. C., & McConnell, S. R. (2014). Redefining individual growth and development indicators: Oral language. *Assessment for Effective Intervention, 39,* 233–244. doi:10.1177/1534508413496837

Burns, M. K., Appleton, J. J., & Stehouwer, J. D. (2005). Meta-analytic review of responsiveness-to-intervention research: Examining field-based and research-implemented models. *Journal of Psychoeducational Assessment, 23,* 381–394. doi:10.1177/1534508413496837

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96,* 31–42.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology, 96,* 671–681.

Catts, H., Fey, M., Tomblin, J. B., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research, 45,* 1142–1157.

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

Coyne, M. D., McCoach, D. B., & Kapp, S. (2007). Vocabulary intervention for kindergarten students: Comparing extended instruction to embedded instruction and incidental exposure. *Learning Disability Quarterly, 30,* 74–88. doi:10.2307/30035543

Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli, R., Ruby, M., Crevecoeur, Y. C., & Kapp, S. (2010). Direct and extended vocabulary instruction in kindergarten: Investigating transfer effects. *Journal of Research on Educational Effectiveness, 3,* 93–120.

Dickinson, D. (2011, August 19). Teachers' language practices and academic outcomes of preschool children. *Science, 333,* 964–967. doi:10.1126/science.1204526

Dickinson, D., & Porche, M. (2011). Relation between language experiences in preschool classrooms and children's kindergarten and fourth-grade language and reading abilities. *Child Development, 82,* 870–886. doi:10.1111/j.1467-8624.2011.01576.x

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test–Fourth Edition*. Minneapolis, MN: Pearson Assessments.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness, 2*(1), 1–44. doi:10.1080/19345740802539200

Foorman, B., & Torgesen, J. (2001). Critical elements of classroom and small-group instruction promote reading success in all children. *Learning Disabilities Research & Practice, 16,* 203–212. doi:10.1111/0938-8982.00020

Gersten, R., Compton, D., Connor, C., Dimino, J., Santoro, L. E., Linan-Thompson, S., & Tilly, W. (2008). *Assisting students struggling with reading: Response to intervention and multi-tier intervention for reading in primary grades. A practice guide*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Science, U.S. Department of Education.

Gettinger, M., & Stoiber, K. (2008). Applying a response-to-intervention model for early literacy development in low-income children. *Topics in Early Childhood Special Education, 27,* 198–213.

Gray, S., & Brinkley, S. (2011). Fast mapping and word learning by preschoolers with specific language impairment in a supported learning context: Effect of encoding cues, phonotactic probability, and object familiarity. *Journal of Speech, Language, and Hearing Research, 54,* 870–884.

Greenwood, C. R., Carta, J. J., Atwater, J., Goldstein, H., Kaminski, R., & McConnell, S. (2013). Is a response to intervention (RTI) approach to preschool language and early literacy instruction needed? *Topics in Early Childhood Special Education, 33,* 48–64. doi:10.1177/0271121412455438

Greenwood, C. R., Carta, J. J., Kelley, E. S., Guerrero, G., Kong, N. Y., Atwater, J., & Goldstein, H. (in press). Systematic replication of the effects of a supplementary, technology-assisted, storybook intervention for preschool children with weak vocabulary and comprehension skills. *The Elementary School Journal*.

Hadley, P., Simmerman, A., Long, M., & Luna, M. (2000). Facilitating language development for inner-city children: Experimental evaluation of a collaborative, classroom-based intervention. *Language, Speech, and Hearing Services in Schools, 31,* 280–295.

Hargrave, A., & Sénéchal, M. (2000). A book reading intervention with preschool children who have limited vocabularies: The benefits of regular reading and dialogic reading. *Early Childhood Research Quarterly, 15,* 75–90.

Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experiences of young American children*. Baltimore, MD: Brookes.

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. doi:10.3102/0162373707299706

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology, 49,* 4–14. doi:10.1037/a0027238

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2,* 88–110. doi:10.1080/19345740802539325

Johnson, C., & Yeates, E. (2006). Evidence-based vocabulary instruction for elementary students via storybook reading. *EBP Briefs, 1*(3), 1–23.

Justice, L. M., Mashburn, A., Hamre, B., & Pianta, R. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early Childhood Research Quarterly, 23,* 51–68. doi:10.1016/j.ecresq.2007.09.004

Justice, L. M., Meier, J., & Walpole, S. (2005). Learning new words from storybooks: An efficacy study with at-risk kindergartners. *Language, Speech, and Hearing Services in Schools, 36,* 17–32. doi:10.1044/0161-1461(2005/003)

Kelley, E. S., & Goldstein, H. (2014). Building a Tier 2 intervention: A glimpse behind the data. *Journal of Early Intervention, 36,* 292–312. doi:10.1177/1053815115581657

Kelley, E. S., Goldstein, H., Spencer, T., & Sherman, A. (2015). Effects of automated Tier 2 storybook intervention on vocabulary and comprehension learning in preschool children with limited oral language skills. *Early Childhood Research Quarterly, 31,* 47–61. doi:10.1016/j.ecresq.2014.12.004

Leung, C. B. (2008). Preschoolers' acquisition of scientific vocabulary through repeated read-aloud events, retellings, and hands-on science activities. *Reading Psychology, 29,* 165–193.

Loftus, S., Coyne, M. D., McCoach, D. B., & Zipoli, R. (2010). Effects of a supplemental vocabulary intervention on the word knowledge of kindergarten students at risk for language and literacy difficulties. *Learning Disabilities Research & Practice, 25,* 124–136. doi:10.1111/j.1540-5826.2010.00310.x

Lonigan, C. J., Anthony, J. L., Bloomfield, B. G., Dyer, S. M., & Samwel, C. S. (1999). Effects of two shared-reading interventions on emergent literacy skills of at-risk preschoolers. *Journal of Early Intervention, 22,* 306–322. doi:10.1177/105381519902200406

Lonigan, C. J., & Whitehurst, G. J. (1998). Relative efficacy of parent and teacher involvement in a shared-reading intervention for preschool children from low-income backgrounds. *Early Childhood Research Quarterly, 13,* 263–290.

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research, 80,* 300–335.

Mol, S. E., Bus, A. G., & de Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research, 79,* 979–1007. doi:10.3102/0034654309332561

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4,* 133–142. doi:10.1111/j.2041-210x.2012.00261.x

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

National Reading Panel. (2000). *Report of the National Reading Panel: Reports of the subgroups.* (NIH Pub. No. 00-4754). Washington, DC: National Institute of Child Health and Human Development Clearinghouse.

Nelson, K. E., Welsh, J. A., Trup, E. M. V., & Greenberg, M. T. (2011). Language delays of impoverished preschool children in relation to early academic and emotion recognition skills. *First Language, 31,* 164–194. doi:10.1177/0142723710391887

Neuman, S. B. (1999). Books make a difference: A study of access to literacy. *Reading Research Quarterly, 34,* 286–311.

Neuman, S. B., Newman, E. H., & Dwyer, J. (2011). Educational effects of a vocabulary intervention on preschoolers' word knowledge and conceptual development: A cluster-randomized trial. *Reading Research Quarterly, 46,* 103–129. doi:10.1598/RRQ.46.3.3

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research, 78,* 33–84. doi:10.3102/0034654307313793

Pas, E. T., & Bradshaw, C. P. (2012). Examining the association between implementation and outcomes. *The Journal of Behavioral Health Services & Research, 39,* 417–433. doi:10.1007/s11414-012-9290-2

Pence, K. L., Justice, L. M., & Wiggins, A. K. (2008). Preschool teachers' fidelity in implementing a comprehensive language-rich curriculum. *Language, Speech, and Hearing Services in Schools, 39,* 329–341. doi:10.1044/0161-1461(2008/031)

Penno, J. F., Wilkinson, I. A. G., & Moore, D. W. (2002). Vocabulary acquisition from teacher explanation and repeated listening to stories: Do they overcome the Matthew effect? *Journal of Educational Psychology, 94,* 23–33. doi:10.1037//0022-0663.94.1.23

Pullen, P. C., Tuckwiller, E. D., Konold, T. R., Maynard, K. L., & Coyne, M. D. (2010). A tiered intervention model for early vocabulary instruction: The effects of tiered instruction for young students at risk for reading disability. *Learning Disabilities Research & Practice, 25,* 110–123.

Qi, C., Kaiser, A., Milan, S., & Hancock, T. (2006). Language performance of low-income African American and European American preschool children on the PPVT-III. *Language, Speech, and Hearing Services in Schools, 37,* 5–16. doi:10.1044/0161-1461(2006/002)

Rhoads, C. H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics, 36,* 76–104.

Silverman, R. (2007). A comparison of three methods of vocabulary instruction during read-alouds in kindergarten. *Elementary School Journal, 108,* 97–113.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Snow, C., Burns, M., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Research Council.

Spencer, E. J., Goldstein, H., & Kaminski, R. (2012). Teaching vocabulary in storybooks: Embedding explicit vocabulary instruction for young children. *Young Exceptional Children, 15,* 18–32.

Spencer, E. J., Goldstein, H., Sherman, A., Noe, S., Tabbah, R., Ziolkowski, R., & Schneider, N. (2012). Effects of an automated vocabulary and comprehension intervention: An early efficacy study. *Journal of Early Intervention, 34,* 195–221. doi:10.1177/1053815112471990

Spencer, T. D., Goldstein, H., Kelley, E. S., Sherman, A., & McCune, L. (2015). *A curriculum-based measure of language comprehension for preschoolers: Reliability and validity of the Assessment of Story Comprehension (ASC)*. Manuscript submitted for publication.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38,* 934–947. doi:10.1037/0012-1649.38.6.934

Storkel, H. L., & Hoover, J. R. (2011). The influence of part-word phonotactic probability/neighborhood density on word learning by preschool children varying in expressive vocabulary. *Journal of Child Language, 38,* 628–643. doi:10.1017/S0305000910000176

VanDerHeyden, A., Snyder, P., Broussard, C., & Ramsdell, K. (2008). Measuring response to early literacy intervention with preschoolers at risk. *Topics in Early Childhood Special Education, 27,* 232–249. doi:10.1177/0271121407311240

VanDerHeyden, A., Witt, J., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a response to intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45,* 225–256.

van Kleeck, A. (2006). Fostering inferential language during book sharing with prereaders: A foundation for later text comprehension strategies. In A. van Kleeck (Ed.), *Sharing books and stories to promote language and literacy* (pp. 269–318). San Diego, CA: Plural Publishing.

van Kleeck, A. (2008). Providing preschool foundations for later reading comprehension: The importance of and ideas for targeting inferencing in storybook-sharing interventions. *Psychology in the Schools, 45,* 627–643.

van Kleeck, A., van der Woude, J., & Hammett, L. (2006). Fostering literal and inferential skills in Head Start preschoolers with language impairments using scripted book-sharing discussions. *American Journal of Speech-Language Pathology, 15,* 85–94.

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child Development, 65,* 606–621. doi:10.2307/1131404

Wasik, B. A., 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. doi:10.1037/0022-0663.98.1.63

Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical Evaluation of Language Fundamentals Preschool–Second Edition.* San Antonio, TX: Pearson Assessments.

Zucker, T. A., Solari, E. J., Landry, S. H., & Swank, P. R. (2013). Effects of a brief tiered language intervention for prekindergartners at risk. *Early Education & Development, 24,* 366–392. doi:10.1080/10409289.2012.664763

Appendix 1.

Sample Scripts for Embedded Vocabulary and Story Question Lessons

Lesson	Sample script	Notes
Embedded vocabulary	*Wow! The Jungle Friends are thrilled! They are excited to go to the carnival. Thrilled. Say thrilled.* [pause] *Thrilled means excited. Tell me, what word means excited?* [pause] *Thrilled! Good work! When are you thrilled?* [pause] *What about when you get a present. Or your friends come over to play? I bet that makes you feel excited. Now lift the flap.* [Picture of two young boys with party hats and a birthday cake] *Look! These boys are at a birthday party. They are thrilled! They are excited. Tell me, what does thrilled mean?* [pause] *Excited! That's right*.	In each embedded vocabulary lesson, children were provided multiple opportunities to respond, indicated above by pauses in the script. Response opportunities always included repetition of the word, saying the word in response to the definition, and providing the definition. In some lessons, children could lift a flap to see a picture related to the target word. In other lessons, children had the opportunity to provide a gesture or facial expression related to the target word. Positive feedback was provided by the narrator for many of the response opportunities. Because the scripts were prerecorded, all children received the same opportunities to respond as well as the same positive feedback.
Embedded story question	*Do you think Pablo Porcupine's friends will listen to him the next time they play together?* [pause] *Yes, because they learned that Pablo could help them! Pablo was quiet, but he knew how to get to the carnival. I bet they'll listen to Pablo from now on!*	In each embedded story question lesson, children were provided with a single opportunity to respond, indicated above by pauses in the script. After the response opportunity, the narrator provided a model of an appropriate answer as well as a brief explanation of the answer.

Note. Sample scripts are in press by Paul H. Brookes Publishing Co. and are printed with permission of the publisher. Copyright © Paul H. Brookes Publishing Co.

Appendix 2.

Intervention and Assessment Timeline

Week	Story	Measures
Series 1: *Jungle Friends*		
1	Introductory: *Meet the Jungle Friends*	UVT 1 pretest and ASC *Jungle Friends* concept word pretest
2	*Ellie's First Day*	Listening center observations
3	*Leo's Brave Face*	Listening center observations
4	*Jungle Friends Go to the Beach*	Listening center observations
5	Review: *Jungle Friends Bake a Cake*	UVT 1 posttest UVT 2 pretest and ASC
6	*Marquez Monkeys Around*	Listening center observations
7	*If Elephants Could Fly*	Listening center observations
8	*Leo Loses His Roar*	Listening center observations
9	Review: *Jungle Friends Go to the Park*	UVT 2 posttest UVT 3 pretest and ASC
10	*Ellie Gets Stuck*	Listening center observations
11	*A New Jungle Friend*	Listening center observations
12	*Marquez's Backwards Day*	Listening center observations
13	Review: *Jungle Friends Swinging Through the Vines*	UVT 3 posttest and ASC *Jungle Friends* concept word posttest
Series 2: *Forest Friends*		
14	Introductory: *Meet the Forest Friends*	UVT 4 pretest *Forest Friends* concept word pretest
15	*Pablo's Prickly Problem*	Listening center observations
16	*Suki's Sleepover Surprise*	Listening center observations
17	*Bobby's EmBEARassing Visit*	Listening center observations
18	Review: *Suki Squirrel Goes Swimming*	UVT 4 posttest UVT 5 pretest and ASC
19	*Fae's Nose Knows*	Listening center observations
20	*Snow Day for Fae*	Listening center observations
21	*Where is Bobby Bear?*	Listening center observations
22	Review: *Pablo Packs a Picnic*	UVT 5 posttest UVT 6 pretest and ASC
23	*Pablo's Map Matters*	Listening center observations
24	*Fae's Smelly Situation*	Listening center observations
25	*Suki's Selfish Saturday*	Listening center observations
26	Review: *Forest Friends Go to the Library*	UVT 6 posttest and ASC *Forest Friends* concept word posttest

Note. Introductory and review books are labeled; all other books are instructional books. Children listened to instructional books three times and the review book once. Children listened to books in small-group listening centers in their classrooms. Assessments were administered individually by research staff. UVT = Unit Vocabulary Test; ASC = Assessment of Story Comprehension.