



Effects of a Supplemental Vocabulary Program on Word Knowledge and Passage Comprehension

Helen Apthorp , Bruce Randel , Trudy Cherasaro , Tedra Clark , Margaret McKeown & Isabel Beck

To cite this article: Helen Apthorp , Bruce Randel , Trudy Cherasaro , Tedra Clark , Margaret McKeown & Isabel Beck (2012) Effects of a Supplemental Vocabulary Program on Word Knowledge and Passage Comprehension, Journal of Research on Educational Effectiveness, 5:2, 160-188, DOI: [10.1080/19345747.2012.660240](https://doi.org/10.1080/19345747.2012.660240)

To link to this article: <https://doi.org/10.1080/19345747.2012.660240>



Published online: 03 Apr 2012.



Submit your article to this journal [↗](#)



Article views: 989



View related articles [↗](#)



Citing articles: 31 View citing articles [↗](#)

Effects of a Supplemental Vocabulary Program on Word Knowledge and Passage Comprehension

Helen Apthorp

McREL, Denver, Colorado, USA

Bruce Randel

Century Analytics, Centennial, Colorado, USA

Trudy Cherasaro and Tedra Clark

McREL, Denver, Colorado, USA

Margaret McKeown and Isabel Beck

University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Abstract: A cluster randomized trial estimated the effects of a supplemental vocabulary program, Elements of Reading[®]: vocabulary on student vocabulary and passage comprehension in moderate- to high-poverty elementary schools. Forty-four schools participated over a period spanning 2 consecutive school years. At baseline, 1,057 teachers and 16,471 students from kindergarten, first, third, and fourth grade participated. The schools were randomly assigned to either the primary or intermediate grade treatment group. In each group, the nontreatment classrooms provided the control condition. Treatment classrooms used the intervention to supplement their core reading program, whereas control classrooms taught vocabulary business-as-usual. The intervention includes structured, weekly lesson plans for 6 to 8 literary words and aural/oral and written language activities providing multiple exposures and opportunity for use. Hierarchical linear modeling was used to estimate both proximal (Year 1) and distal (Year 2) effects on vocabulary and passage comprehension. The intervention had positive and statistically significant proximal effects but no statistically significant distal effects. The results indicate that the intervention can improve targeted vocabulary and local passage comprehension, but expecting global effects may be overly optimistic.

Keywords: Vocabulary, elementary school, intervention, efficacy

The benefits of a rich vocabulary for young children accrue rapidly as they become readers. Correlational research suggests that children's oral vocabulary in preschool, kindergarten, and first grade predicts future reading achievement in both later elementary grades and high school (Cunningham & Stanovich, 1997; Muter, Hulme, Snowling & Stevenson, 2004; Storch & Whitehurst, 2002).

It is the oral vocabularies of emergent readers that first allow them access to meaning when applying letter-sound correspondences (National Reading Panel, 2000). With continued reading practice and developing fluency, the reading experience itself builds vocabulary (Cunningham & Stanovich, 1993; Wagner, Muse, & Tannenbaum, 2007).

Address correspondence to Helen Apthorp, Marzano Research Laboratory, 9000 East Nichols Avenue, Suite 112, Englewood, CO 80112, USA. E-mail: helen.apthorp@marzanoresearch.com

Cromley and Azevedo's (2007) research modeling the comprehension process demonstrates the key role of vocabulary in reading for students in intermediate to secondary grades. Their model included background knowledge, inferences, comprehension strategies, vocabulary, and word reading in ninth graders. They found that vocabulary and background knowledge made the greatest contributions to comprehension. The reciprocal relationship between vocabulary and comprehension is likely to be at least part of the motivation for researchers to target instruction in vocabulary as a direction for improving reading comprehension.

Researchers have suggested that features of multiple encounters and active processing of target words in vocabulary instruction may lead to comprehension gains through mechanisms such as increased fluency of access to word meanings while reading (Beck, Perfetti, & McKeown, 1982; Mezynski, 1983; Nation & Snowling, 1999; Ouelette, 2006; Perfetti, 2007) and greater activation of connections to concepts associated with the target words (Landauer, 2009; Perfetti & Adlof, *in press*). Silverman's (2007) study suggests a refinement in how "active processing" is conceived; finding that instruction requiring analytic responses to word meaning was more effective than instruction that included asking students to talk about their own experiences related to words that were taught. Decades of research also suggests that successful vocabulary instruction provides learners both definitional and contextual information about target words, in addition to the features of multiple encounters and active processing (Baumann, Kame'enui, & Ash, 2003; Beck, McKeown, & Kucan, 2002; Blachowicz & Fisher, 2000; Mezynski, 1983; Nagy & Scott, 2000; National Reading Panel, 2000; Rupley & Nichols, 2005; Stahl & Fairbanks, 1986). More recent instructional studies have confirmed the importance of these features with learners as young as preschool (Collins, 2010; Coyne et al., 2010; Silverman, 2007).

Current meta-analyses confirm the effectiveness of vocabulary instruction for improving vocabulary, and with less magnitude, improving comprehension (Elleman, Lindo, Morphy & Compton, 2009; Marulis & Newman, 2010). Based on experimental studies only, the average effect sizes range from 0.29 to 1.21 for the impact on vocabulary and from 0.10 to 0.50 for the impact on comprehension (Elleman et al., 2009; Marulis & Newman, 2010). Although correlational, the results of moderator analyses were consistent with the research-based recommendations mentioned earlier, suggesting that successful vocabulary instruction includes both explicit provision of definitions and opportunities for children to derive word meaning experientially in context while reading or listening to connected text (Marulis & Newman, 2010). Other moderator analyses indicated that successful vocabulary instruction for targeted objectives (such as understanding a discrete set of words or skills) is brief—with a duration of a week or less (Elleman et al., 2009; Marulis & Newman, 2010). In addition, more discussion in vocabulary instruction was associated with greater success in learning and using new vocabulary (Elleman et al., 2009).

Many researchers contend that vocabulary interventions need to begin early and continue for a sustained period in order to make a substantial impact (Beck & McKeown, 2007; Biemiller, 2003; Foorman, Seals, Anthony, & Pollard-Durodola, 2003). The need for sustained, effective vocabulary instruction is especially great among children whose lives outside of school provide limited exposure to rich language—especially text-based language—and vocabulary. Difference in language exposure between children in welfare families and the children in professional families is thought to be a major contributor to differences in young children's vocabulary repertoires, measured in thousands of words (Hart & Risley, 1995).

THE INTERVENTION

The instructional design of *Elements of Reading®: Vocabulary* (EOR: V; Beck & McKeown, 2004), based on the research just summarized and intended for teachers' daily classroom use across consecutive grades, K to 5, offers an opportunity to test the efficacy of early and sustained vocabulary instruction. In addition to incorporating both definitional and contextual information, multiple exposures to and use of new words, and structured activities to engage children's deep and active processing, the developers of EOR: V targeted Tier Two words. Tier Two words characterize written language, rather than everyday spoken language (Beck et al., 2002). Recommendations about word selection for vocabulary instruction vary, but most vocabulary researchers agree that the words to be taught should be ones that will be important to success in children's future learning (Blachowicz, Fisher, & Watts-Taffe, 2006). Stahl and Nagy (2006) referred to these kinds of words as "high utility general vocabulary" and described them as "words you'd expect to be part of a literate person's vocabulary" (p. 61). Their examples for the first nine letters of the alphabet are *abolish*, *banish*, *chamber*, *deliberate*, *exceed*, *frequent*, *genuine*, *hospitable*, and *initiative*. The words themselves provide more sophisticated or refined labels for familiar concepts; for example, *chamber* is used in written language for "small room"; *genuine* is a more sophisticated way of expressing "real"; and *benevolent*, another Tier Two word, can be understood if one knows "kind" (Beck et al., 2002).

Depending on grade level, EOR: V introduces six to eight Tier Two words per lesson over 24 lessons per grade. The lessons each follow a three-phase structure to introduce the vocabulary, use the vocabulary, and assess the vocabulary in a 5-day/weekly lesson plan intended for delivery in 10 to 20 min daily. On Day 1, teachers introduce words with both contextual and definitional information through the use of read-alouds, student-friendly explanation, and photo cards with dramatic images to help children visualize and personalize each word. On Days 2, 3, and 4 of the lesson, teachers guide students through the use of words in different activities that prompt the children to think about and apply their knowledge of the new words to everyday situations and episodes. It is these teacher-guided *bringing words to life* activities that provide children multiple exposures and opportunities to use words. On Day 5, teachers guide students in a vocabulary review and assessment; the teacher provides cumulative review opportunities in subsequent lessons.

PURPOSE

Although the program was in widespread use during and prior to Reading First (Florida Center for Reading Research, 2003; Educational Research Institute of America, 2005), to date, however, the efficacy of the program had not been evaluated in a controlled randomized trial. The primary purpose of this study is to provide unbiased estimates of the average effect of EOR: V on vocabulary and passage comprehension in schools serving children from low-income households. The theory underlying the EOR: V intervention holds that EOR: V's instantiation of the three components of successful vocabulary interventions identified in the literature (i.e., definitional and contextual information about word meaning, multiple exposures, and deep processing) would lead to greater average scores on two proximal outcomes for the treatment group relative to the control group: (a) learning of target Tier Two words, and (b) ability to comprehend passages using the target words. Through frequent encounters and thoughtful activities with target words, it is theorized that EOR: V will enable students to develop increased speed and/or efficiency of semantic access

to well-connected networks of words and meaning that will be helpful when reading or listening to passages for comprehension. We hypothesized that continued exposure to EOR: V for 2 consecutive years would produce measurable changes in student performance on distal measures of standardized tests of vocabulary and passage comprehension.

To test these hypotheses, we collected data in the context of a cluster randomized trial, with participating schools randomly assigned to one of two groups. We evaluated two proximal effects (increased vocabulary and passage comprehension using target words) and two distal effects (increased performance on standardized tests of vocabulary and passage comprehension). We conducted the experiment for two school years, 2008–2009 and 2009–2010, and children continued from one grade to the next over the consecutive 2-year period in the same experimental condition (EOR: V or control).

METHOD

Participants and Random Assignment

Forty-six elementary schools in six school districts in one southeastern state participated in this study. Three districts were members of a cooperative services consortium in a rural area; the other three districts included two large and one midsize suburban school district according to the 2008/09 NCES locale variable. The study used a four-block design classifying the three rural districts into one block and the remaining three districts into individual blocks.

Schools within each block were randomly assigned to either Group A or B. Group A schools used EOR: V in primary grade classrooms (K, 1, and 2) and provided a control condition in the intermediate grades (3, 4, and 5), whereas Group B schools provided the opposite set of conditions (i.e., used EOR: V in intermediate grades and provided a control condition in the primary grades). The assignment of schools to group created two tandem studies occurring simultaneously: the Primary and Intermediate Grade studies. It was deemed that there was little risk of “treatment spillage” (i.e., control-group members picking up and using parts of the intervention) because the intervention and its materials are very specific to each grade level. After random assignment, two schools withdrew from Group B, leaving 44 schools in the final study sample (one school withdrew before the baseline assessment and one school withdrew afterwards). Table 1 shows the baseline distribution of participating schools by block and group.

At baseline, the teacher sample included 1,057 teachers and the student sample included 16,471 students. Comparison of treatment and control groups at baseline on student achievement, controlling for nesting of students within schools, revealed no statistically

Table 1. Number of participating schools by block and group at baseline

Block	Group A	Group B
Rural Cooperative Services Consortium	4	4
Large District 1	8	9
Large District 2	8	8
Medium-size District	2	2
Total	22	23

significant difference on SAT-10 at baseline. In the Primary Grade Study ($N = 45$ schools), the baseline SAT-10 treatment and control group standards score averages of 525.34 and 533.09, respectively, had a difference of -7.75 ($SE = 4.00$, $p = .06$). In the Intermediate Grade Study ($N = 45$ schools), the baseline SAT-10 treatment and control group averages of 600.95 and 593.83, respectively, had a difference of 7.11 ($SE = 5.274$, $p = .18$). There were no statistically significant differences between groups of schools in total enrollment and percentage of free or reduced-price lunch and English-language learners students. The final sample of 44 schools had an average enrollment of 632 students, with an average of 75% students eligible for free or reduced-price lunch, and 20% English-language learners.

Orientation, Training, and Timeline

Orientation

District Reading/Language Arts coordinators served as block-level study coordinators, and at each participating school, assistant principals coordinated data collection and reading coaches trained and supported treatment teachers. During the Year 1 study orientation with these key district and school staff, researchers explained the study design and distributed re-search information (e.g., a study timeline, and list of roles and responsibilities). Participants also received written notification of the group to which their school was assigned.

Training

The research team conducted district-level EOR: V training workshops with the reading coaches from each participating school, including distribution of the EOR: V training package (PowerPoint, video clips), demonstration of how to use the training package, and small-group discussion about EOR: V implementation. Subsequently, reading coaches scheduled and conducted EOR: V training with treatment teachers at their school. Treatment teachers were asked to supplement their core/basal reading program by implementing EOR: V using the 5-day, weekly lesson plans and materials provided in the Teacher's Guide and classroom kit. Control teachers did not receive any extra training in vocabulary instruction from their reading coaches. Control teachers were asked to implement their reading/language arts vocabulary instruction in a business-as-usual manner.

Timeline

In Year 1 (school year 2008–2009), researchers provided the study orientation and EOR: V training to a rural consortium of schools in October 2008 and to the other districts in December 2008. The EOR: V classroom kits and materials for treatment teachers arrived at consortium schools in early November 2008 and at the other district schools in January 2009.¹ The baseline student achievement test was administered between October 2008 and January 2009, and the Year 1 posttest was administered in May 2009 to all kindergarten, and first-, third-, and fourth-grade classrooms. Student participants matriculated between Year 1 and Year 2 to the first, second, fourth, and fifth grades. For Year 2 (school year 2009–2010), in September 2009, a refresher study orientation and EOR: V training was

¹The late start for three districts was due to withdrawal of a district and 20 schools in August 2008 necessitating recruitment of replacement schools during fall 2008.

provided and EOR: V classroom materials arrived at all participating schools. The Year 2 student posttest was administered in March 2010. Site visits and classroom observations were conducted in December and February in Year 1 and in November in Year 2.

Measurement

Student Achievement

All student assessments were group administered by trained examiners who were not teachers of the students being assessed.

Prior Achievement. To control for prior academic performance, treatment and control students were administered a SAT-10 subtest prior to the start of the intervention in Year 1. The Listening to Words and Stories subtest was administered to kindergarten and first-grade classrooms and the Reading Vocabulary subtest was administered to third- and fourth-grade classrooms. The reliability estimates for the subtests are high enough to distinguish groups of students (frequently in the 0.80s, but range from 0.53 to 0.93; Carney, 2005; Morse, 2005).

Vocabulary and Passage Comprehension Year 1. For the Year 1 posttest, we administered two researcher-developed assessments designed for this study, the *Tests of Instructed Word Knowledge* in vocabulary (TOIW-V) and comprehension (TOIW-C). Each grade level of the TOIW-V and TOIW-C assessed students' knowledge of Tier Two words randomly sampled from the first 15 lessons of EOR: V.²

The TOIW-V was an orally group-administered instrument developed specifically to assess recognition and understanding of EOR: V words. The TOIW-V is based on a verbal task developed by two members of the research team (Beck & McKeown, 2007) and used by Coyne (Coyne, McCoach, Loftus, Zipoli, & Kapp, 2009) that involves answering yes or no to questions such as, "Does *exhausted* mean to borrow something?"

The TOIW-C was developed to test students' ability to use their knowledge of target words contained in passages to understand what was happening and why (Beck & McKeown, 2007). The test passages were primarily narratives; for example, two children want to do something *humane* to help families who lost their belongings in a fire. The test at each level comprised four passages, each using five target words from the first 15 lessons of EOR: V program. The kindergarten and first grade TOIW-C was administered orally with each passage read-aloud twice. The students answered questions by selecting one of three pictures they had viewed while the question was asked. The third- and fourth-grade TOIW-C was administered as a paper-and-pencil assessment; students read passages and answered written multiple-choice questions. The questions at all levels included a variety of "what," "when," and "why" questions about the events and people in the passages.

We conducted separate item analyses on TOIW-V and TOIW-C at each grade level. A minimum point-biserial correlation, $pb(r) < .10$, was used to eliminate items that showed weak associations with overall scores, except for the Grade-K and Grade-1 comprehension assessments, where all items were retained due to the small number of items. Reliability

²The TOIW words were randomly selected from the first 15 lessons rather than the entire 24-lesson EOR: V program to match implementation fidelity expected with the late start in Year 1 (January)—which was completion of at least the first 15 lessons before the end of the school year.

Table 2. Student counts, scores and reliability coefficients for Year 1 posttests

Grade	TOIW-V				TOIW-C			
	<i>N</i>	<i>M (SD)</i>	No. of Items/ Points Possible	Reliability Coefficient	<i>N</i>	<i>M (SD)</i>	No. of Items/ Points Possible	Reliability Coefficient
K	1,531	45.79 (9.24)	80	.77	1,588	9.81 (2.80)	20	.41
1	17,777	52.54 (9.69)	80	.84	1,668	9.33 (2.44)	20	.46
3	1,572	57.65 (9.14)	78	.84	1,744	10.75 (3.64)	20	.72
4	1,581	68.00 (11.11)	96	.87	1,450	12.43 (4.03)	24	.75

coefficients (Cronbach's alpha) and number of items used to calculate the composite scores for each instrument are presented in Table 2. Reliability coefficients for all instruments ranged from 0.41 to 0.87. Lowest estimates were found on the Grade-K and Grade-1 TOIW-C instruments. To account for the error in the Grade-K and Grade-1 TOIW-C measures, we estimated effects of EOR: V on these outcomes by correcting for attenuation (Hunter & Schmidt, 1990).

Year 1 Student Sampling. We randomly assigned classrooms to take one or the other TOIW-V or TOIW-C, because the entire student sample was not needed for sufficient statistical power. First, we ranked classrooms within each school and grade from lowest to highest based on median baseline SAT-10 scores. We randomly assigned the classroom with the lowest median score to take one or the other (TOIW-V or TOIW-C) using a coin toss ("tails" was assigned to the vocabulary test). We then moved up the list, alternating assignment to TOIW-V and TOIW-C based on that initial random assignment. We repeated the same process for each grade in a school.

Vocabulary and Passage Comprehension Year 2. For the Year 2 posttest, we administered selected subtests of The Group Reading and Diagnostic Evaluation (GRADE; Williams, 2001), a group-administered, norm-referenced test of reading and reading-related skills. The GRADE was selected as the Year 2 outcome assessment because it included measures of vocabulary and listening and vocabulary and reading passage comprehension for the primary and intermediate grades, respectively. Reliability estimates range from 0.76 to 0.95. Validity evidence suggests that the GRADE is an appropriate measure of the identified student outcomes as evidenced by concurrent and predictive validity studies and analysis of convergent and divergent validity (Williams, 2001).

We administered the GRADE word meaning and listening comprehension subtests in the primary grades. GRADE word meaning asks students to read a word and choose a picture that best tells the meaning of the word. GRADE listening comprehension requires students to listen to one or more sentences and then identify the picture that best matches the sentence(s) to assess vocabulary, grammar, inference-making, and idiom and nonliteral comprehension.

We administered the GRADE reading vocabulary and passage comprehension subtests in the intermediate grades. GRADE reading vocabulary requires students to read a

word and identify a synonym. GRADE passage comprehension requires students to read both synthetic and authentic passages on various topics and answer multiple choice questions that assess four comprehension strategies—questioning, clarifying, summarizing, and predicting.

Year 2 Student Sampling. Because we did not require Year 2 posttest data from all students for sufficient statistical power to estimate the impact of EOR: V, we randomly selected three classes at each grade within each school to take the two GRADE subtests. In schools with fewer than three classes in a grade, all classes took the two GRADE subtests.

Teacher Instruments

Teacher Survey. Teachers completed a one-time background survey asking about years of teaching experience and the identity of their core/basal reading program. Each study year, teachers completed a survey of vocabulary instruction where they recorded the total number of instructional days available for a specified 1-week period, and the number of minutes spent in direct vocabulary instruction each day of that week. On the survey, direct vocabulary instruction was defined as “intentional, planned instruction in word meaning and word use for the purposes of expanding, refining or deepening student vocabulary knowledge.”

The survey also asked teachers to report how many days in that same week they used *each* of a variety of instructional activities (e.g., provided word definitions; asked students to identify a picture that showed word meaning; asked students to explain why they thought a picture showed the meaning of a word). Each activity item score was recoded by dividing the number of days for the activity by the total number of instructional days available to represent the proportion of available time used for the activity. Composite scores were computed for four types of activities: active processing (e.g., asked students to explain their word choice), deep processing (e.g., students explained why they thought a picture showed the meaning of a word), passive responding (e.g., directly taught definitions), and associative processing (e.g., students identified synonyms or antonyms). The alpha coefficients were .84, .80, .81, and .79 for the active, deep, passive, and associative composite scores, respectively.

Recorded Vocabulary Instruction Lessons. In Year 1, teachers audio-recorded a vocabulary lesson implemented between April 20 and May 1 following step-by-step directions. We asked both treatment and control teachers to select a large group 10- to 15-min vocabulary lesson involving question and answering so that verbal interaction between teacher and students could be recorded. The transcribed audio-recorded lessons were coded to identify and categorize teacher questions about words into one of four types: recall/repeat, connect, generate, and integrate/explain.

Vocabulary Lesson Observations. In both Year 1 and 2, researchers conducted onsite classroom observations during vocabulary lessons. The Year 1 EOR: V observation checklist included a list of EOR: V lesson activities. Observers recorded “yes/no” based on whether the activity was observed. The Year 2 EOR: V checklist expanded upon the Year 1 checklist by including items that could be counted (e.g., number of friendly definitions provided by the teacher). In Year 2, we also used a checklist for observations of control classroom

vocabulary instruction, including vocabulary activities not particular to EOR: V (e.g., teachers modeled how to use context to infer word meaning).

Treatment Teacher Logs. Each year, treatment teachers completed EOR: V lesson logs indicating the extent to which they used each of 10 activities in the EOR: V 5-day lesson plan as directed in the Teacher's Guide. For some activities, teachers rated the extent to which they performed the activity on a 4-point scale, from 0 (*not at all*) to 1 (*hardly at all*) to 2 (*to some extent*) to 3 (*to a great extent*), for other activities they recorded a "yes/no" response corresponding to whether or not they performed an activity, and for still other activities the response was a count. In addition, the lesson logs asked teachers to identify the EOR: V lessons they had used. In Year 1, EOR: V lesson logs were administered weekly; in Year 2, they were administered quarterly. In Year 2, teachers identified their most recently used EOR: V lesson and reported the extent to which they used each lesson activity in that particular lesson. Composite scores were computed for offering contextual information (read-alouds) and definitional information (providing friendly explanations), providing multiple exposures (percentage of activities across the 5-day lesson sequences implemented), and encouraging deep processing (extent to which teachers requested student explanations).

Teacher Sampling. All treatment and control teachers were asked to complete the one-time background survey and the annual survey of vocabulary instruction. In Year 1, we randomly selected one teacher per grade level per school to audio-record a vocabulary lesson. In both years, all treatment teachers were asked to complete lesson logs. In Year 1, we selected one treatment classroom to observe. In Year 2, we selected 20 centrally located schools to visit (10 Group A and 10 Group B schools) and randomly selected one second-grade teacher and one fourth-grade teacher per school to observe.

Fidelity of Implementation

The core reading programs were divided evenly among three top-selling basals in both Group A and B schools. Reading coaches reported satisfaction with the EOR: V structure (e.g., "user-friendly," "program is very well scripted—helps teachers plan") and materials (e.g., "visual [photo] cards give students more opportunities to apply the words") but concerns about the amount of time to implement, particularly for completing some of the read-aloud selections ("I think it takes 30 minutes not 10 to 20 minutes daily").

Dosage and Instructional Time

Dosage was measured by whether the EOR: V lesson that the teacher was observed using was the intended lesson for that week based on the EOR: V Teacher's Guide, a one-lesson-per-week assumption, and the school's start-up date. In Year 1, between 43 and 65% of teachers at each grade level demonstrated appropriate dosage. In Year 2, between 45 and 67% of teachers at each grade level demonstrated appropriate dosage (see Table 3).

According to teacher self-report, Year 1 teachers typically implemented at least 15 of the EOR: V lessons, which was the intended number of lessons given the late start in Year 1. According to teacher self-report, the Year 2 median number of lessons used out of 24 possible was 24 in first and second grade and 23 and 20 in fourth and fifth grade, respectively. Also according to Year 2 self-report, the median number of words per weekly

Table 3. Observed EOR: V dosage and procedural fidelity

	Year 1 (2008/09)				Year 2 (2009/10)	
	K ^a	First ^b	Third ^c	Fourth ^d	First ^e	Fourth ^f
Met EOR: V lesson dosage criterion ^g (%)	43	52	65	45	67	45
Proportion of EOR: V lessons that met criteria						
General activities ^h						
<i>M</i>	.90	.88	.90	.89	.70	.64
(<i>SD</i>)	(.12)	(.16)	(.17)	(.14)	(.23)	(.19)
Day-specific activities ⁱ						
<i>M</i>	.85	.87	.91	.89	.59	.80
(<i>SD</i>)	(.18)	(.27)	(.24)	(.19)	(.37)	(.25)
Overall procedural (general plus specific)						
<i>M</i>	.88	.88	.90	.89	.63	.73
(<i>SD</i>)	(.11)	(.19)	(.13)	(.11)	(.27)	(.19)

Note. *N* is the count of teachers.

^a*N* = 21. ^b*N* = 23. ^c*N* = 20. ^d*N* = 22. ^e*N* = 12. ^f*N* = 11. ^gDosage was measured by whether the Elements of Reading: Vocabulary (EOR: V) lesson that the teacher was observed using was the intended lesson for that week based on the EOR: V Teacher's Guide, a one lesson per week assumption, and the school's start-up date for that study year. ^hA sample general EOR: V lesson activity is *Teacher refers to Teacher's Guide or demonstrates proficiency with its content*. ⁱSample activities specific to the day observed in the 5-day EOR: V Lesson sequence include: *Teacher orally reads the Read-aloud to class* (Day 1); *Teacher leads the class through creation of a Word Organizer* (Day 4).

lesson for which teachers provided EOR: V friendly explanations of meaning was equal to the intended number.

The Year 1 and 2 survey results indicated that use of EOR: V took additional instructional time, which is consistent with the intended supplemental nature of EOR: V. In both years, treatment teachers reported more daily minutes of vocabulary instruction than control teachers. In Year 1, the EOR: V teachers spent 3 to 11 min more daily than the control teachers in vocabulary instruction. In Year 2, EOR: V teachers spent zero to 3 more min daily than control teachers in vocabulary instruction. Differences were statistically significant (see Table 4). Therefore, we included instructional time as a covariate in the impact analysis models to account for the effect of time on students' posttest performance.

Procedural Fidelity

Year 1 observations indicated that EOR: V teachers, on average, demonstrated high levels of procedural fidelity, implementing 85 to 91% of the intended activities (see Table 3). Teacher self-report in the Year 1 EOR: V lesson logs were consistent with these observations. For read-alouds, use of photo cards, and student practice books, 82 to 100% of log reports met fidelity criteria. For more complex and open-ended activities, such as use of word chats, writing, and asking students "why" follow-up questions or to explain their reasoning, between 61 and 79% of reports met fidelity criteria. Among fourth-grade teachers, however, procedural fidelity was higher: 93% of log reports met fidelity criteria.

Table 4. Comparison of instructional time for vocabulary in treatment and control

	No. of Minutes of Daily Vocabulary Instruction				Estimated Difference
	Treatment Group		Control Group		
	<i>N</i> Teachers	<i>M</i>	<i>N</i> Teachers	<i>M</i>	
<hr/>					
Year 1					
Kindergarten	96	15.74	101	12.37	3.37* (1.10)
Grade 1	113	15.69	111	12.09	3.60** (0.98)
Grade 3	114	17.96	106	11.68	6.28** (0.95)
Grade 4	88	22.17	85	11.08	11.10** (1.03)
Year 2					
Primary grades	194	11.82	195	11.73	0.09 (0.93)
Intermediate grades	146	14.35	151	11.23	3.12*** (0.82)

Note. Vocabulary instruction was defined as “intentional, planned instruction in word meaning and word use for the purposes of expanding, refining or deepening student vocabulary knowledge.” Numbers in parentheses are standard errors.

* $p < .05$. ** $p < .01$.

Fourth-grade treatment teachers explained that they used the EOR: V writing prompts to prepare students for the state assessment.

Year 2 observations indicated that EOR: V teachers, on average, implemented 59 to 73% of the intended activities (see Table 3). Year 2 EOR: V teacher self-report yielded estimates consistent with the observations for providing definitional and contextual information. The lesson log data, however, indicated low procedural fidelity for providing multiple exposure and use opportunities and engaging students in deep processing. Based on Year 2 self-report, 24 to 50% of teachers met the fidelity criteria for multiple exposures and depth of processing.

Quality of Instruction

Year 1 lesson transcript analysis revealed qualitative differences in vocabulary instruction in treatment and control classrooms (see Table 5). Of the 74 requests for lesson audio-recordings from teachers, 56 (76%) were usable and transcribed (30 from treatment classrooms and 26 from control classrooms). Analyses revealed that in control classroom vocabulary lessons, 50% of the teachers' questions were recall/repeat questions and 50% higher order questions (connect, generate, and integrate/explain), whereas in EOR: V vocabulary lessons, 20% of the teachers' questions were recall/repeat and 80% were higher order with an emphasis on connect questions. The EOR: V teachers, guided by the EOR: V lesson plans, asked students to make comparisons that connected a picture or context that was associated with the word, for example, “Which is *delicate*, a bike helmet or a small glass vase?” In contrast, left on their own, control teachers asked students to generate

Table 5. Year 1 comparison of quality of vocabulary instruction in treatment and control lessons: distribution of teacher questions by type

	Total Coded Questions	Question Type			
		Read/Repeat and Recall	Connect	Generate	Integrate and Explain
Control ^a	610	307 (50.5%)	97 (16%)	180 (30%)	26 (4.3%)
Treatment ^b	1,046	213 (20%)	467 (44.6%)	230 (22%)	136 (12.6%)

Note. *Read/repeat:* Teacher asked students to read, repeat, or copy in writing a word, definition, or sentence containing a taught word. *Recall:* Teacher asked students to recall the word given the definition, or recall the definition, given the word. *Connect:* Teacher asked students to make a connection between a word and a picture or context; for example, “Which is *delicate*, a bike helmet or a small glass vase?” *Generate:* Teacher asked students to generate their own examples of the word or an antonym or synonym. *Integrate:* Teacher asked students to respond to a context that contained a target word in a way that required integrating the word’s meaning into the context; for example, “How could you help someone who felt *distraught*?” *Explain reasoning:* Teacher asked students to explain the reasoning behind their response, such as “Why would you be *curious* if you heard lots of laughter from the classroom next door?” (McKeown, Beck, & Apthorp, 2011).

^a*N* = 26. ^b*N* = 30.

examples, but the examples were often repetitive and not potentially useful to clarifying or deepening understanding of a word’s meaning (McKeown, Beck & Apthorp, 2011). For example, a control teacher asked, “What famous person would you like to have a *conversation* with?” Students named their choices, such as John Legend, Will Smith, and so on, with eight students taking turns; but in the generation of examples, the word *conversation*, or the idea of talking to these people, never came up.

Also, based on self-report in survey results, treatment teachers spent proportionally more of their vocabulary instructional time encouraging deep processing and proportionally less time engaging students in passive responses (see Table 6). In Year 2, treatment teachers reported using between 40 and 46% of their vocabulary instructional time for deep-processing activities compared with control teachers reporting between 33 and 34% of their vocabulary instructional time for deep-processing activities (see Table 6).

Targeted vocabulary also differed in EOR-V and control vocabulary lessons. Seventy-six percent of the 21 Year 2 observations of control teachers’ vocabulary lessons emphasized content domain words (e.g., survive, habitat, concentrate) and Tier One words (i.e., high-frequency words). Five control lessons (24%) emphasized Tier Two words (e.g., *drudgery*, *smoldering*, *dither*). Instructional activities varied in control classroom vocabulary lessons; variations were providing friendly explanations of word meaning, morphological study of words, modeling comprehension strategies, modeling word pronunciation and word use, and different student activities (e.g., word games on the computer; students generating sentences).

To summarize implementation fidelity, each study year, up to two thirds of EOR: V teachers implemented the intended dosage when measured by number of lessons used. In Year 2, self-reported number of words taught in terms of providing friendly explanations of word meaning met adequate fidelity criteria in each grade level. Procedural fidelity was inconsistent within the treatment condition (e.g., ranging from 24 to 93% meeting adequate

Table 6. Year 2 comparison of quality of vocabulary instruction in treatment and control lessons: proportion of time spent on different activities

	Proportion of Time Spent on Different Instructional Activities				Estimated Difference
	Treatment Group		Control Group		
	<i>N</i> Teachers	<i>M</i>	<i>N</i> Teachers	<i>M</i>	
Primary grades					
Active response ^a	186	.48	192	.43	.04 (.02)
Deep processing ^b	185	.46	193	.34	.10*** (.03)
Passive response ^c	184	.51	191	.57	−.06* (.02)
Associative response ^d	183	.46	192	.46	−.00 (.02)
Inter-mediate grades					
Active response	139	.44	150	.41	.02 (.03)
Deep processing	143	.40	150	.33	.06* (.03)
Passive response	141	.48	149	.51	−.03 (.02)
Associative response	142	.42	150	.42	.00 (.02)

Note. Numbers in parentheses are standard errors.

^aActive processing, for example: asked students to explain their word choice. ^bDeep processing, for example: students explained why they thought a picture showed the meaning of a word. ^cPassive responding, for example: directly taught definitions. ^dAssociative processing, for example: students identified synonyms or antonyms.

* $p < .05$. *** $p < .001$.

fidelity criteria on multiple encounters and active processing of target words) but distinct from control classroom vocabulary instruction in both quality and amount of time. EOR: V teachers spent more time daily than control teachers in vocabulary instruction and asked proportionally more questions that connected word meaning to contexts, whereas control teachers asked proportionally more questions that required recall or repetition. Also, EOR: V teachers reported spending proportionally more time than control teachers eliciting deep processing.

Data Analysis

To estimate the impact of EOR: V on Year 1 student outcomes, we conducted three-level hierarchical linear model analyses where students were nested within classrooms and classrooms nested within schools. The models included an individual-level pretest covariate (classroom mean centered), a classroom-level pretest covariate (school mean centered) and a school-level pretest covariate (grand mean centered) using the SAT-10 baseline performance. The school level of the model also included the treatment indicator

and grand mean centered dummy variables to account for the blocks used in random assignment. Models were run separately at each grade level for the vocabulary and passage comprehension outcomes in Year 1. To account for instructional time as a factor, we subsequently reran the Year 1 impact analysis models including instructional time as a classroom-level covariate (school mean centered).

The impact of EOR: V in Year 2 was estimated using a two-level model with students nested in schools. Because students had been in two different classrooms by the end of Year 2, the classroom level was not used in the impact estimate. However, as noted in Moerbeek (2004), failing to model the middle level of a multilevel model still results in correct estimates of the standard error for fixed effects. Because the two primary grades (first and second) and the two intermediate grades (fourth and fifth) were administered the same standardized measures of vocabulary and passage comprehension, the grades were combined within level in each impact analysis. The Year 2 models included pretest covariates at the individual (school mean centered) and school levels (grand mean centered), a grand mean centered dummy variable for grade at the individual level to control for the proportion of students in each grade across the analysis sample, and grand mean centered dummy variables at the school level to account for the blocks used in random assignment. Data attrition and baseline equivalence of all impact analysis samples were evaluated.

RESULTS

Proximal Effects of EOR: V in Year 1

At the end of Year 1, EOR: V had a statistically significant effect on each Year 1 outcome except for Grade 3 TOIW Comprehension (see Table 7). To account for instructional time, a second model was evaluated that included instructional time as a covariate; there were no substantive changes in the statistical significance or the magnitude of the EOR: V effects (see the appendix). In both models, with and without instructional time, the seven statistically significant effects of EOR: V remained significant at the .05 level after correcting for multiple comparisons with a Benjamini–Hochberg correction (What Works Clearinghouse, 2008). The effect sizes from both models were similar, with the magnitude of effects both increasing and decreasing when instructional time was included as a covariate by no more than 19 tenths of an effect size. Therefore, patterns of results are discussed in relation to the original impact analysis reported in Table 7.

For the vocabulary outcomes, the EOR: V effects were stable and large across the four grades. The effect sizes for TOIW Vocabulary ranged from 0.95 in third grade to 1.24 in fourth grade. For passage comprehension (TOIW-C), the EOR: V effect sizes were smaller in magnitude, ranging from a low of 0.21 and 0.24 for TOIW-C Kindergarten and TOIW-C first grade (after correcting for low reliability of the outcome measure) to a high of 0.44 for TOIW-C in fourth grade (see Glass's *d*, bottom row, Table 7).

Evaluating Possible Bias

Because data attrition could create selection bias in the estimated proximal effects of EOR: V, we evaluated overall and differential attrition and checked baseline equivalence in the impact analysis samples. The overall data attrition ranged from 18 to 23% with differential

Table 7. Impact of EOR: V on Year 1 student outcomes without instructional time covariate

Outcome	TOIW-V K	TOIW-V Grade 1	TOIW-V Grade 3	TOIW-V Grade 4	TOIW-C K	TOIW-C Grade 1	TOIW-C Grade 3	TOIW-C4 Grade 4
Fixed effects								
Intercept	45.41**** (0.85)	49.53**** (0.71)	53.34**** (0.64)	57.74**** (0.98)	10.19**** (0.22)	9.2**** (0.12)	10.33**** (0.19)	11.31**** (0.27)
EOR: V	6.98**** (0.72)	7.85**** (1.01)	7.11**** (0.88)	10.87**** (0.91)	1.02**** (0.21)	0.75**** (0.15)	0.33 (0.27)	1.63**** (0.38)
Level 1 covariate	0.04**** (0.01)	0.05**** (0.01)	0.06**** (0.01)	0.65**** (0.01)	0.01**** (0.01)	0.01**** (0.01)	0.02**** (0.01)	0.04**** (0.01)
Level 2 covariate	0.06**** (0.01)	0.05**** (0.01)	0.02**** (0.01)	0.10**** (0.02)	0.02**** (0.01)	0.02**** (0.01)	0.01** (0.01)	0.01 (0.01)
Level 3 covariate	0.10**** (0.03)	0.08* (0.04)	0.07* (0.03)	0.10** (0.03)	0.03**** (0.01)	0.03**** (0.01)	0.02* (0.01)	0.06**** (0.01)

Block 1	-0.68 (0.92)	-1.55 (1.30)	-3.56 (1.11)	-0.06 (1.18)	-0.34 (0.27)	-0.37 (0.21)	-0.35 (0.36)	-0.94 (0.55)
Block 2	0.26 (1.32)	1.87 (2.02)	-0.41 (1.74)	0.54 (1.78)	0.77 (0.43)	-0.05 (0.21)	0.96 (0.54)	0.40 (0.74)
Block 3	-0.94 (1.13)	-0.10 (1.57)	-0.60 (1.41)	1.55 (1.42)	0.41 (0.32)	-0.60 (0.25)	0.08 (0.41)	-0.41 (0.59)
Random effects								
Residual	58.33****	58.52****	44.59****	64.18****	6.51****	4.87****	10.25****	9.74****
Classroom	6.13****	9.57****	6.51****	10.88****	0.29*	0.33**	1.51****	2.44****
School	1.17	5.71*	3.95*	1.98	0.16	0.01	—	—
Teacher ICC	0.09	0.13	0.12	0.14	0.04	0.06	0.13	0.20
School ICC	0.02	0.08	0.07	0.03	0.02	0.00	0.00	0.00
Glass's d	0.98	1.00	0.95	1.24	0.38 ^a	0.31 ^a	0.09	0.44

Note. Teacher intraclass correlation (ICC) is the ratio of the teacher level variance component to the total variance. School ICC is the ratio of the school level variance component to the total variance. Glass's *d* is estimated as the ratio of the treatment effect estimate to the estimate of the total variance in the control group estimated from an unconditional model (not shown). Numbers in parentheses are standard errors. EOR: V = Elements of Reading; Vocabulary; K = kindergarten. TOIW-V = Test of Instructed Words-Vocabulary; TOIW-C = Test of Instructed Words-Comprehension.

^aAfter correcting for attenuation due to low reliability of the outcome measure, the TOIW-C Kindergarten effect size is 0.24 and the TOIW-C first-grade effect size is 0.21.

* $p < .05$. ** $p < .01$. *** $p < .001$. **** $p < .0001$; all two-tailed.

attrition between 2 and 4% (see Table 8). These are allowable levels with minimal risks for selection bias (What Works Clearinghouse, 2008). Baseline achievement (SAT-10) was not significantly different except for the first-grade vocabulary impact analysis sample (see Table 9). These differences were accounted for by including the pretest covariate (SAT-10) in the analysis model.

We also evaluated overall and differential attrition in the impact analysis that included instructional time as a covariate. The overall attrition in these analyses samples ranged from 32 to 35%, and the differential attrition ranged from 1 to 9% (see the appendix, Table A2). The differential attrition rates in the third- and fourth-grade samples (8% and 9%, respectively) in combination with the overall attrition of 32 and 34 were above the allowable rates, raising concerns about possible selection bias. We examined two possible sources of bias in these impact analysis samples: teacher experience and prior achievement (SAT-10). With regards to teachers' experience, we identified moderately experienced teachers versus new teachers (defined as having 4 to 10 years of experience vs. 1 to 3 years of experience) because of their differential impact on student achievement (Darling-Hammond, 2000; Rivkin, Hanushek, & Kain, 2005). In the present analysis samples, there were no significant differences in the distribution of moderately experienced and new teachers. The treatment and control impact analysis samples each had approximately 28% of teachers with 1 to 3 years experience, 40% with 4 to 10 years, and 33% with 11 or more years. Checking baseline equivalence on prior achievement revealed that there were no significant differences between treatment and control except in the first grade TOIW Vocabulary impact analysis sample (see the appendix, Table A3). We controlled for this potential bias by including prior achievement (SAT-10) as a pretest covariate. Therefore, the potential for selection bias was mitigated by demonstrating equivalence in teacher experience, baseline equivalence, and inclusion of a prior achievement covariate.

Distal Effects of EOR: V in Year 2

In Year 2, EOR: V had no statistically significant impact on the standardized measures of vocabulary or passage comprehension in either the primary or intermediate grades (see Table 10). Distributions of the Year 2 outcome scores were examined at each grade level, and no ceiling effects were evident. The null effects indicated that as measured, EOR: V did not produce distal effects (see Tables 10 and 11).

DISCUSSION

In this study, effects of a widely used supplemental vocabulary program were evaluated in a cluster randomized trial involving moderate to high-poverty elementary schools and a 10-month intervention spanning 2 consecutive school years. We evaluated both proximal and distal effects. Proximal effects were evaluated by measuring student understanding of the words targeted by the program. Distal effects were evaluated by measuring student vocabulary and passage comprehension with a standardized test.

Table 7. Year 1 and 2 student data attrition

	Total			Treatment			Control		
	N	N	Attrition (%)	N	Tested N	Attrition (%)	N	Tested N	Attrition (%)
Year 1									
Kindergarten TOIW	3,844	3,119	23	1,719	1,411	19	2,125	1,708	20
First-grade TOIW	4,211	3,445	18	2,068	1,671	19	2,143	1,774	17
Third-grade TOIW	4,166	3,316	20	2,079	1,701	18	2,087	1,615	23
Fourth-grade TOIW	3,839	3,031	21	1,923	1,481	23	1,916	1,550	19
Year 2									
Primary grades word meaning	4,516	2,618	42	2,250	1,146	49	2,266	1,472	35
Primary grades listening comprehension	4,516	2,803	38	2,250	1,326	41	2,266	1,477	35
Intermediate grades reading vocabulary	5,469	3,492	36	2,670	1,730	35	2,670	1,762	34
Intermediate grades passage comprehension	5,469	3,478	36	2,670	1,925	28	2,670	1,753	34

Note. In Year 1, the sample size represents the number of all students with a SAT-10 score. Year 1 tested represents the number of students in the impact analysis sample. These were students with a SAT-10 score who also had a score available on the Test of Instructed Words–Vocabulary (TOIW-V) or the Test of Instructed Words–Comprehension (TOIW-C). In Year 2, the sample size represents the number of students in classes randomly selected to take the Group Reading Assessment and Diagnostic Evaluation (GRADE), see the previous description of sampling procedures. Year 2 tested represents the number of students in the impact analysis sample. These were students with a SAT-10 score who also had a score available on the GRADE.

Table 8. Comparison of baseline achievement for impact analysis samples

Outcome	Treatment Group <i>M</i>	Control Group <i>M</i>	Estimated Difference
Year 1			
TOIW-V Kindergarten	517.75	516.73	1.02 (5.90)
TOIW-V Grade 1	573.27	554.37	18.90* (7.24)
TOIW-V Grade 3	568.76	582.81	−14.05 (12.58)
TOIW-V Grade 4	607.74	608.06	−0.32 (6.82)
TOIW-C Kindergarten	508.34	510.50	−2.16 (7.43)
TOIW-C Grade 1	562.46	552.64	9.82 (5.13)
TOIW-C Grade 3	567.56	581.36	−13.80 (11.51)
TOIW-C Grade 4	580.45	597.81	−17.36 (16.20)
Year 2			
Primary grades	530.71	536.21	−5.50 (4.03)
Intermediate grades	605.31	596.93	8.38 (5.68)

Note. Numbers in parentheses are standard errors. TOIW-V = Test of Instructed Words–Vocabulary; TOIW-C = Test of Instructed Words–Comprehension.

* $p < .05$.

Proximal Effects

At the end of Year 1 and after 4 months of program implementation, statistically significant and positive effects of Elements of Reading: Vocabulary were found for kindergarten, first-, third-, and fourth-grade students on custom measures of vocabulary. The Year 1 effect sizes, ranging from 0.95 to 1.24, all were larger than 0.79, an estimated average effect size for the impact of vocabulary interventions on custom measures of vocabulary from a recent meta-analysis (Elleman et al., 2009). Also at the end of Year 1 in three of the four grades, the program had statistically significant and positive effects on student comprehension of passages containing the words that were taught. EOR: V produced effect sizes of 0.21, 0.24, and 0.44 for custom measures of passage comprehension in kindergarten, first-, and fourth-grade, respectively.

Effect sizes, as standard deviation units, can be translated into percentile point gains or losses on a set of normally distributed scores for the average person in the comparison group if he or she had received the intervention (What Works Clearinghouse, 2011). For the Year 1 outcomes, all effects were positive which means the average student in the comparison group, which by definition has a 50th percentile rank score, would gain in percentile rank if he or she received EOR: V. For example, for the largest and smallest effect sizes reported above for TOIW-V (1.24) and TOIW-C (0.21), the difference in performance would be 39

Table 9. Impact of EOR: V on Year 2 outcomes

Outcome	Primary Word Meaning	Intermediate Vocabulary	Primary Listening Comprehension	Intermediate Passage Comprehension
Fixed effects				
Intercept	24.73**** (0.15)	97.68**** (1.20)	15.44**** (0.07)	16.84**** (0.47)
EOR: V	0.22 (0.21)	−1.92 (1.70)	0.09 (0.10)	−0.66 (0.67)
Level 1 covariate	0.01**** (0.01)	0.06**** (0.01)	0.01**** (0.01)	0.02**** (0.01)
Level 2 covariate	0.03*** (0.01)	0.14* (0.06)	0.01** (0.01)	0.05 (0.02)
Grade dummy	0.95**** (0.15)	−1.65**** (0.42)	−0.11 (0.07)	0.10 (0.19)
Block 1 dummy	0.81 (0.37)	−2.70 (3.16)	0.32 (0.17)	−0.30 (1.23)
Block 2 dummy	0.78 (0.31)	1.38 (2.44)	0.08 (0.13)	0.94 (0.95)
Block 3 dummy	0.57 (0.27)	1.39 (2.16)	−0.06 (0.12)	0.32 (0.84)
Random effects				
Residual	12.66****	150.80****	2.69****	31.32****
School	0.25**	29.01****	0.05**	4.24****
ICC within schools	0.98	0.84	0.98	0.88
ICC between schools	0.02	0.16	0.02	0.12
Glass’s <i>d</i>	0.06	−0.14	0.05	−0.11

Note. Numbers in parentheses are standard errors. Effect sizes were calculated using Glass’s *d* (Glass, McGaw, & Smith 1981). All outcomes are subtests of the Group Reading Assessment and Diagnostic Evaluation (American Guidance Service, 2001). EOR: V = Elements of Reading: Vocabulary; ICC = intraclass correlation.

p* < .01. **p* < .0001; all two-tailed.

Table 10. Mean scores for Year 2 posttest (group reading assessment and diagnostic evaluation subtests) by grade level

	Primary		Intermediate	
	Listening Comprehension	Word Meaning	Vocabulary ^a	Passage Comprehension
N	2,803	2,618	3,492	3,478
<i>M</i>	15.26	24.52	98.54	17.21
<i>SD</i>	1.70	3.71	14.40	6.24
Highest possible score	17	27	145	28

Note. The Primary grade level includes first- and second-grade students, and the Intermediate grade level includes fourth- and fifth-grade students.

^aVocabulary subtest scores represent scaled scores.

and 8 percentile rank points for the average fourth grade and kindergarten control students, respectively, if they received EOR: V.

The addition of instructional time as a covariate in the impact analysis models did not change the statistical significance or the magnitude of the estimated proximal effects of EOR: V. Moreover, the observation and survey results indicated unique qualitative patterns to the additional vocabulary instruction provided by treatment teachers. Treatment teachers targeted Tier Two words, used proportionally more deep processing activities, and asked more higher order questions.

The lack of a statistically significant effect of EOR: V on third-grade comprehension of passages containing the instructed words ($ES = .09$) may have been a function of the reading demands of the third-grade assessment. The Year 1 third-grade outcome test required students to read passages and answer multiple-choice questions. Even though EOR: V students may have been familiar with the vocabulary included in the stories, the program that taught those words emphasized their use in aural/oral activities. There may not have been sufficient practice independently reading the target words in sentences or passages or sufficient practice answering written multiple-choice questions for students to successfully meet the reading demands of the proximal comprehension measure.

Distal Effects

In addition to the proximal effects, we hypothesized that with lexical growth in Year 1, and continued participation in EOR: V in the subsequent year, we would be able to detect distal effects on standardized measures. As reported in the results section for Year 2, no statistically significant distal effects of EOR: V were detected.

The EOR: V students' lexical growth, evident in the Year 1 results, and which may have continued to be positive in Year 2, did not transfer to, or was not evident as, a positive influence on students' ability to perform on the Year 2 standardized measures of vocabulary and comprehension. The small Year 1 effects on passage comprehension (although statistically significant) in combination with the null effects in Year 2 have implications for our original theory of action. We predicted that frequent encounters and thoughtful activities with Tier Two words would enable students to develop fast access to semantically accurate and well-connected networks of words and meaning. We also predicted that over a sustained time in EOR: V, students would develop more fluent and deeper verbal reasoning in general, and this would serve as a mechanism in improving passage comprehension. The present results suggest that the impact of EOR: V may be limited to local comprehension and that expecting broader impacts may be overly optimistic.

To determine whether broader impacts of EOR: V is overly optimistic, we need a better understanding of the contribution of vocabulary to comprehension and assessments capable of measuring those contributions. Identifying gains on standardized comprehension tests from vocabulary interventions is rare. For example, the effects for standardized measures in the Elleman et al. (2009) meta-analysis in contrast to experimenter-designed measures were negligible. However, these researchers and others note that standardized measures may not be sensitive to the contribution of vocabulary to comprehension (Elleman et al., 2009; Pearson, Hiebert, & Kamil, 2007).

The present study provided evidence that EOR: V leads to comprehension of short passages containing the instructed words. It may be that the EOR: V intervention promoted gains in comprehension that are not measureable by traditional standardized assessments. Or

it may be that such gains did not occur, but the instruction setup conditions for future gains, for example, by leading students to more readily learn unfamiliar words that will accumulate over time to sufficiently affect comprehension. This possibility would be consistent with Landauer's view; based on his work with Latent Semantic Analysis (Landauer & Dumais, 1997), every word learned reduces the number of sentences a reader will encounter with unknown words (Landauer, 2009).

Yet another possibility is that to affect comprehension, students need more language input, beyond that available in vocabulary lessons, in which to place their new knowledge and practice its integration into reading. Thus, a major issue at play here is that we have not yet discerned the mechanisms by which vocabulary development leads to comprehension. Understanding such mechanisms would seem to go hand in hand with developing assessments capable of measuring the true effects of vocabulary interventions.

Another perspective is that "individual differences in the richness of conceptual knowledge" is the contributing factor associated with vocabulary that influences passage comprehension (Wagner et al., 2007, p. 286). Perfetti (2007) used the notion of lexical quality to encompass both conceptual richness and efficiency of vocabulary knowledge to explain differences in passage comprehension associated with vocabulary knowledge. Lexical quality supports the "central connecting event between word identification and text comprehension" (Perfetti, 2007, p. 375) which is when readers/listeners identify a word and integrate its meaning into their ongoing representation of the input. Suggesting that the source of lexical quality is knowledge, both conceptual and linguistic, variation can arise through literacy and language experiences, including, "among other things, learning to decode printed words, practice reading and writing, and engagement with concepts and their language forms" (Perfetti, 2007, p. 380).

If to affect comprehension students need more language input beyond that available in vocabulary lessons, then the next step in research and development is determine how to leverage the proximal effects of EOR: V with additional literacy and language experiences.

Limitations

In this study, moderate to strong proximal effects were estimated in a randomized control trial of a supplemental vocabulary program. The design of the trial, however, was not without its flaws. Although instructional time was refuted as a confound in impact analyses that included time as a covariate, and by evidence of qualitative differences between treatment and control classroom instruction, the limitation of the correlational and post hoc nature of controlling time is important to acknowledge. In addition, teacher-related factors also included in post hoc analyses (e.g., teaching experience) remain viable causal factors. Future research on the efficacy of supplemental vocabulary instruction is needed that controls the time and teacher-related factors as part of the initial design.

Another limitation is that the proximal effects were detected on measures tightly aligned with the content and procedures of the instructional program. Other measures representing different processing demands, genres, and content knowledge are needed to evaluate less proximal effects of EOR: V. Because the proximal measures used primarily narrative texts, future research would be helpful on evaluating the effects of EOR: V on comprehension of informational text.

Implications for Practice and Research

In the present study, educators who used EOR: V appreciated the premade explanations of word meaning, photo card representations of word meaning, and multiple opportunities for student practice through varied teacher prompts and guided activities. However, difficulties arose in managing the activities within the expected time. The present results and discussion suggest it important for prospective users of EOR: V to be clear on expectations and carefully weigh options for the focus of teachers' limited instructional time. First, as a supplemental intervention, the expectation is that EOR V be used briefly, in 10- to 20-min lessons daily. Educators need to understand that extending that time risks the reduction of instructional time for different and additional kinds of language and literacy input, content, and experiences. Second, in making resource adoption decisions, it will be important to carefully consider if EOR: V's targeted words are the right words; if they are not, EOR: V probably is not a good choice.

At the same time, we know that EOR: V worked to some extent for promoting children's vocabulary across multiple grades in typical elementary school settings. Future research and development is needed addressing how to refine EOR: V so teachers are likely to use it more efficiently. The questions to address concern how to provide contextual information and multiple opportunities for exposure and application more efficiently (e.g., Should the read-aloud passages be shortened, eliminated, or selected using different criteria? Are word chat prompts and other teacher questions effective when asked of students during transitions between activities, or assigned for homework or as a computer-assisted activity?). Also, how might the lesson structure and design apply to the development of instruction for other types or sets of words? Are there generic protocols and sentence frames that can be used to develop lesson scripts for teacher-use to promote vocabulary across the content areas using a common approach?

The other questions to address in future research are about the role of vocabulary in comprehension. Present results do not support the view that EOR: V, as implemented, is likely to bolster comprehension of material that does not contain the targeted words. Although comprehension depends on understanding individual words, reading for understanding requires other important skills and knowledge, including broad conceptual knowledge (both general and discipline-specific knowledge), grammar, word-level skills, and the ability to make inferences, self-monitor and apply mental effort (Shanahan et al., 2010; Snowling & Hulme, 2011). It may be that for vocabulary instruction to affect comprehension, it needs to be more tightly integrated with rather than supplementary to these aspects of comprehension.

REFERENCES

- Baumann, J. F., Kame'enui, E. J., & Ash, G. E. (2003). Research on vocabulary instruction: Voltaire redux. In J. Flood, D. Lapp, J. R. Squire, & J. M. Jensen (Eds.), *Handbook of research on teaching the English language arts* (pp. 752–785). Mahwah, NJ: Erlbaum.
- Beck, I. L., & McKeown, M. G. (2004). *Elements of reading vocabulary: Teacher's guide level A* (Item No. 0-7398-8458-1). Austin, TX: Steck-Vaughn.
- 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.
- Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York, NY: Guilford.

- Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, 74, 506–521.
- Biemiller, A. (2003). Vocabulary: Needed if more children are to read well. *Reading Psychology*, 24, 323–335.
- Blachowicz, C. L. Z., & Fisher, P. (2000). Vocabulary instruction. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Volume III* (pp. 503–523). Mahwah, NJ: Erlbaum.
- Blachowicz, C. Z., Fisher, P. L., Ogle, D., & Watts-Taffe, S. (2006). Vocabulary: Questions from the classroom. *Reading Research Quarterly*, 41(4), 524–539. doi:10.1598/RRQ.41.4.5
- Carney, R. N. (2005). Review of the Stanford Achievement Test, 10th Edition. In R. A. Spies & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook*. Retrieved from the Buros Institute's Test Reviews Online website: <http://www.unl.edu/buros>
- Collins, M. F. (2010). ELL preschoolers' English vocabulary acquisition for storybook reading. *Early Childhood Research Quarterly*, 25, 84–97.
- Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli, R., & Kapp, S. (2009). Direct vocabulary instruction in kindergarten: Teaching for breadth vs. depth. *Elementary School Journal*, 110, 1–18.
- Coyne, M. D., McCoach, D. B., Loftus, S., Zipoli, R., Ruby, M., Crevecoeur, Y., & Kapp, S. (2010). Direct and extended vocabulary instruction in kindergarten: Investigating transfer effects. *Journal of Research on Educational Effectiveness*, 3, 93–120.
- Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, 99, 311–325.
- Cunningham, A. E., & Stanovich, K. E. (1993). Where does knowledge come from? Specific association between print exposure and information acquisition. *Journal of Educational Psychology*, 85, 211–229.
- Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, 33, 934–945.
- Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, 8, 1–50.
- Educational Research Institute of America. (2005, September). *A review of Elements of Reading: Vocabulary using the consumer's guide to evaluating a core reading program*. Bloomington, IN: Author.
- Elleman, A., Lindo, E., Morphy, P., & Compton, D. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, 2, 1–44.
- Florida Center for Reading Research. (2003, June). *Elements of reading: Vocabulary*. Tallahassee, FL: Author.
- Foorman, B. R., Seals, L. M., Anthony, J., & Pollard-Durodola, S. (2003). A vocabulary enrichment program for third and fourth grade African-American students: Description, implementation, and impact. In B. R. Foorman (Ed.), *Preventing and remediating reading difficulties* (pp. 419–441). Timonium, MD: York.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Thousand Oaks, CA: Sage.
- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. New York, NY: Sage.
- Landauer, T. K. (2009, February 19). *Towards teaching the right words in the right way for the right student at the right time*. Invited seminar, Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 105, 211–240.

- Marulis, L. M., & Newman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research, 80*, 300–335.
- McKeown, M. G., Beck, I. L., & Apthorp, H. (2011, April 10). *Examining depth of processing in vocabulary lessons*. Paper presented at the American Educational Research Association conference, New Orleans, LA.
- Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research, 53*, 253–279.
- Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*, 129–149.
- Morse, D. T. (2005). Review of the Stanford Achievement Test, 10th edition. In R. A. Spies & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook*. Retrieved from the Buros Institute's Test Reviews Online website: <http://www.unl.edu/buros>
- Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology, 40*(5), 666–681.
- Nagy, W. E., & Scott, J. A. (2000). Vocabulary processes. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research: Volume III* (pp. 269–284). Mahwah, NJ: Erlbaum.
- Nation, K., & Snowling, M. J. (1999). Developmental differences in sensitivity to semantic relations among good and poor comprehenders: Evidence from semantic priming. *Cognition, 20*, B1–B13.
- National Reading Panel. (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4754). Washington, DC: National Institute of Child Health and Human Development.
- Ouelette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading and reading comprehension. *Journal of Educational Psychology, 98*, 554–566.
- Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly, 42*, 282–296.
- Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*, 357–383.
- Perfetti, C., & Adlof, S. M. (in press). Reading comprehension: A conceptual framework for word meaning to text meaning. In *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences*. Lanham, MD: Rowman & Littlefield Education.
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417–458.
- Rupley, W. H., & Nichols, W. D. (2005). Vocabulary instruction for the struggling reader. *Reading & Writing Quarterly, 21*, 239–260.
- Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). *Improving reading comprehension in kindergarten through 3rd grade: A practice guide* (NCEE 2010-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <http://whatworks.ed.gov/publications/practiceguides>
- Silverman, R. (2007). A comparison of three methods of vocabulary instruction during read-alouds in kindergarten. *Elementary School Journal, 108*, 97–113.
- Snowling, M. J., & Hulme, C. (2011). Evidence-based interventions for reading and language difficulties: Creating a virtuous circle. *British Journal of Educational Psychology, 81*, 1–23.
- Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*, 72–110.
- Stahl, S. A., & Nagy, W. E. (2006). *Teaching word meanings*. Mahwah, NJ: Erlbaum.
- 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.

- Wagner, R. K., Muse, A. E., & Tannenbaum, K. R. (2007). Promising avenues for better understanding of vocabulary and comprehension. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 276–291). New York, NY: Guilford.
- What Works Clearinghouse. (2008). *Procedures and standards handbook* (Ver. 2.0). Retrieved from <http://ies.ed.gov/ncee/wwc/references/idoctrviewer/doc.aspx?docid=19&tocid=1>
- What Works Clearinghouse. (2011). *Procedures and standards handbook* (Ver. 2.1). Retrieved from <http://ies.ed.gov/ncee/wwc/references/idoctrviewer/doc.aspx?docid=19&tocid=1>
- Williams, K. T. (2001). *Group reading assessment and diagnostic evaluation (GRADE) Technical manual*. Circle Pines, MN: American Guidance Service.

APPENDIX

Table A1. Impact of EOR: V on Year 1 student outcomes with instructional time covariate

Outcome	TOIW-V K	TOIW-V Grade 1	TOIW-V Grade 3	TOIW-V Grade 4	TOIW-C K	TOIW-C Grade 1	TOIW-C Grade 3	TOIW-C Grade 4
Fixed effects								
Intercept	45.87*** (0.98)	50.28*** (0.82)	53.87*** (0.62)	57.78*** (1.08)	10.11*** (0.25)	9.25*** (0.14)	10.23*** (0.23)	11.61*** (0.31)
EOR: V	7.42*** (0.81)	7.16*** (1.19)	6.66*** (0.86)	12.76*** (1.20)	1.02*** (0.24)	0.73*** (0.20)	0.41 (0.32)	1.27* (0.48)
Instructional Time covariate	-0.00 (0.04)	0.14* (0.06)	0.03 (0.05)	-0.19* (0.09)	0.04** (0.02)	0.02 (0.01)	-0.04 (0.03)	0.03 (0.03)
Level 1 pretest covariate	0.05*** (0.01)	0.04*** (0.01)	0.07*** (0.01)	0.06*** (0.01)	0.01*** (0.01)	0.01*** (0.01)	0.02*** (0.01)	0.03*** (0.01)
Level 2 pretest covariate	0.07*** (0.02)	0.05** (0.02)	0.02*** (0.01)	0.09*** (0.02)	0.01*** (0.01)	0.01*** (0.01)	< .01* (0.01)	0.01 (0.01)
Level 3 pretest covariate	0.90** (0.03)	0.08* (0.04)	0.07* (0.03)	0.11** (0.04)	< .01 (0.01)	< .01 (0.01)	< .01 (0.01)	0.08*** (0.02)
Block 1	-0.24 (0.99)	-2.09 (1.45)	-3.70 (1.03)	0.19 (1.35)	-0.40 (0.27)	-0.48 (0.23)	-0.15 (0.40)	-1.18* (0.59)
Block 2	0.23 (1.37)	1.37 (2.17)	-0.49 (1.62)	1.53 (2.04)	0.63 (0.42)	0.09 (0.40)	1.07 (0.66)	-0.28 (0.80)
Block 3	-0.97 (1.24)	-1.16 (1.72)	-0.51 (1.36)	1.62 (1.69)	0.43 (0.33)	-0.87** (0.28)	0.18 (0.44)	-0.26 (0.62)
Random effects								
Residual	46.12*** (2.64)	52.02*** (6.78***)	51.50*** (2.95)	69.68*** (7.94)	6.93*** (0.46)	4.95*** (0.53)	11.22*** (1.53***)	11.50*** (2.21)
Classroom	2.84	1.75	1.57	1.60	0.13	0.14	—	0.40
School	0.06	0.13	0.06	0.11	0.07	0.11	0.16	0.19
Teacher ICC	0.06	0.03	0.03	0.02	0.02	0.03	0.00	0.03
School ICC	1.03	0.92	0.90	1.43	0.37	0.31	0.11	0.34
Glass's <i>d</i>								

Note. Teacher intraclass correlation (ICC) is the ratio of the teacher-level variance component to the total variance. School ICC is the ratio of the school-level variance component to the total variance. Glass's *d* is estimated as the ratio of the treatment effect estimate to the estimate of the total variance in the control group estimated from an unconditional model (not shown). Numbers in parentheses are standard errors. EOR: V = Elements of Reading; Vocabulary; K = kindergarten; TOIW-V = Test of Instructed Words-Vocabulary; TOIW-C = Test of Instructed Words-Comprehension.

^aAfter correcting for attenuation due to low reliability of the outcome measure, the TOIW-C Kindergarten effect size is 0.24 and the TOIW-C first-grade effect size is 0.21.

* $p < .05$. ** $p < .01$. *** $p < .001$. **** $p < .0001$; all two-tailed.

Table A2. Year 1 student data attrition in impact analysis with instructional time covariate

	Total			Treatment			Control		
	Baseline Sample <i>N</i>	Analysis Sample <i>N</i>	Attrition (%)	Baseline Sample <i>N</i>	Analysis Sample <i>N</i>	Attrition (%)	Baseline Sample <i>N</i>	Analysis Sample <i>N</i>	Differential Attrition
Year 1									
Kindergarten TOIW	3,844	2,490	35	1,719	1,158	33	2,125	1,332	4
First-grade TOIW	4,211	2,752	35	2,068	1,357	34	2,143	1,395	1
Third-grade TOIW	4,166	2,835	32	2,079	1,501	28	2,087	1,334	8
Fourth-grade TOIW	3,839	2,517	34	1,923	1,342	30	1,916	1,175	9

Note: The baseline sample size represents the number of all students with a SAT-10 score. The analysis sample size represents the number of students in the impact analysis sample. These were students with a SAT-10 score who also had a score available on the Test of Instructed Words–Vocabulary (TOIW-V) or the Test of Instructed Words–Comprehension (TOIW-C), and had teachers with instructional time data.

Table A3. Comparison of baseline achievement for samples in impact analyses with instructional time covariate

Outcome	Treatment Group <i>M</i>	Control Group <i>M</i>	Estimated Difference
Year 1			
TOIW–V Kindergarten	517.38	515.54	1.84 (4.9)
TOIW–V Grade 1	567.22	553.41	13.81* (6.01)
TOIW–V Grade 3	574.80	585.70	10.89 (10.83)
TOIW–V Grade 4	608.39	608.50	0.11 (6.13)
TOIW–C Kindergarten	509.00	509.48	0.47 (6.14)
TOIW–C Grade 1	559.75	553.03	6.72 (3.81)
TOIW–C Grade 3	573.22	584.95	11.73 (9.72)
TOIW–C Grade 4	586.94	600.99	1405 (14.03)

Note. Numbers in parentheses are standard errors. TOIW–V = Test of Instructed Words–Vocabulary; TOIW–C = Test of Instructed Words—Comprehension.
**p* < .05.