Effects of Academic Vocabulary Instruction for Linguistically Diverse Adolescents: Evidence from a Randomized Field Trial

Nonie K. Lesaux

Harvard Graduate School of Education

Michael J. Kieffer

New York University

Joan Kelley

Julie Russ

Harvard Graduate School of Education

Note
This research was supported by a grant from the Institute of Education Sciences (U.S. Department of Education; Grant No. R305A080631) and a William T. Grant Foundation Scholars Award, both awarded to Nonie K. Lesaux. The authors would like to thank Carol Barry, Teresa Walter, Jennifer Cheatham, Andrea Anushko, Taralynn Kantor, Amy Griffiths, Phoebe Sloane, Mark Nielsen, Tisha Ruibal, Sean Preci, Greta Castaneda, S. Elisabeth Faller, and Armida Lizarraga for their instrumental roles in carrying out the over- all study, as well as David J. Francis and Stephanie M. Jones for their helpful insights. We are also grateful to the participating students and teachers.

Abstract

We conducted a randomized field trial to test an academic vocabulary intervention designed to bolster the language and literacy skills of linguistically diverse sixth-grade students (*N* = 2082; *n* = 1469 from a home where English is not the primary language), many demonstrating low achievement, enrolled in 14 urban middle schools. The 20-week classroom-based intervention improved students' vocabulary knowledge, morphological awareness skills, and comprehension of expository texts that included academic words taught, as well as their performance on a standardized measure of written language skills. The effects were generally larger for students whose primary home language is not English and for those students who began the intervention with underdeveloped vocabulary knowledge.

Keywords: vocabulary, literacy, middle school, curriculum, at-risk students

In this knowledge-based and information-driven global economy, academic success is essential to an individual's life outcomes. Yet many students experience academic failure because of underdeveloped literacy skills (Duncan & Murnane, 2011; Murnane, Sawhill, & Snow, 2012). The growing population of language minority (LM) students, who come from homes where the primary language spoken is not the language of schooling, is at particular risk of school failure. This population is charged with simultaneously developing English language proficiency while also learning academic content, and therefore needs to learn with tremendous efficiency to keep pace with the demands of the curriculum (August & Shanahan, 2006).

Literacy research and instructional initiatives have historically focused on young children, yet there is growing concern about developing evidence-based approaches to promoting adolescents' literacy skills, ensuring their abilities keep pace with what it means to be literate. For example, the vocabulary and language of the young reader's storybook, filled with concrete ideas and objects, is much more straightforward and basic than the abstract language and concepts read by the adolescent studying for an exam (Duke & Carlisle, 2011; Snow & Uccelli, 2009). It has long been understood that early reading difficulties are often exacerbated with increasing grade levels. Now it is becoming clear that young children who fare well in the early grades may struggle later due to the greater complexity of language and content (e.g., Best, Floyd, & McNamara, 2008; Leach, Scarborough, & Rescorla, 2003; Nation, Cocksey, Taylor, & Bishop, 2010; Storch & Whitehurst, 2002). This is especially true for LM students; often they decode and comprehend the conversational language that conveys ideas and topics in beginner books, but lack the sophisticated, abstract vocabulary necessary to support later text comprehension and production (August & Shanahan, 2006; Authors, 2011; Goldenberg, 2011). Moreover, recent research shows this uneven development is also the case for many English-only

(EO) learners (sometimes referred to as *native English-speaking* students) enrolled in high-poverty schools (e.g., Authors, 2010a). Yet, few studies have evaluated specific approaches to advance at-risk adolescents' literacy skills.

Building academic vocabulary—words that appear frequently in texts across academic disciplines, but rarely occur in oral conversation (Baumann & Graves, 2010; Nagy & Townsend, 2012)—is one promising route for improving struggling adolescent learners' academic outcomes. The research reported here tested whether an academic vocabulary intervention, designed for use in middle school classrooms with high proportions of LM students and implemented under typical conditions, would improve language, reading, and writing skills.

Theoretical Foundation for Academic Vocabulary Intervention

Because literacy development is a multifaceted process that demands a number of separate, but related, competencies (Duke & Carlisle, 2011; McCutchen, 2006; RAND Reading Study Group, 2002), there are myriad potential sources of difficulty for the learner who struggles to understand, discuss, and produce academic texts. For middle schoolers, these competencies are largely comprised of higher level processing and linguistic skills. In part, these skills are made up of knowledge that relates to literacy itself: knowledge of process, text-structure, genre, and author (or reader) expectations (Beers & Nagy, 2011; RAND Reading Study Group, 2002; Saddler & Graham, 2007). They also include the ability to draw on prior knowledge, make appropriate inferences, and resolve structural and semantic ambiguities (Alexander & Jetton, 2000;Kintsch & Rawson, 2005). For the learner to undertake this complex process of comprehending and producing academic text, deep and flexible knowledge of the often abstract and complex words and phrases used in this particular register is needed.

Therefore, under-developed language skills have a significant effect on literacy outcomes (Catts, Adlof, & Weismer, 2006; Vellutino, Tunmer, Jaccard, & Chen, 2007). Many struggling adolescents, particularly LM students and their peers growing up in low-income communities, demonstrate under-developed vocabulary knowledge (Authors, 2010a; Buly & Valencia, 2002; Hock et al., 2009). Still, while evidence indicates a strong relationship between academic vocabulary knowledge and literacy development, the degree to which academic vocabulary instruction transfers to broader literacy competencies (e.g., reading comprehension, writing) remains unclear (Baumann, 2009; Elleman, Lindo, Morphy, Compton, 2009; Graham & Perin, 2007; Nagy & Townsend, 2012). In particular, questions remain about how to best target instruction to students' vocabulary learning needs with sufficient intensity, and whether such instruction is likely to increase students' access to academic text and talk.

In response to these questions, we tested the effects of a 20-week intervention, targeting students' knowledge of the specialized academic vocabulary of text, and implemented under typical conditions as part of the core English Language Arts (ELA) instruction in urban middle schools with high numbers of LM students. The intervention design is theoretically grounded in principles of effective vocabulary instruction; principles written about extensively (Baumann, Kame'enui, & Ash, 2003; National Reading Panel, 2000; Stahl & Fairbanks, 1986; Stahl & Nagy, 2006), though rarely operationalized and tested for adolescents from diverse backgrounds.

The first design principle that guided our work is that such instruction must be text-based, so that academic vocabulary words are studied in the authentic contexts in which they are used (Stahl & Fairbanks, 1986; Stahl & Nagy, 2006). Second, in light of the heterogeneous nature of words (for a discussion see Nagy & Hiebert, 2010; Nagy & Scott, 2000), we focused on a particularly high-utility and abstract population of words— general academic words. In this way, each word was necessarily taught from a number of angles, including specialized meanings, its use in various contexts, and its morphological and semantic relationships to other words (Baumann et al., 2002; Baumann & Edwards et al., 2003; Stahl & Nagy, 2006). As a third, related design principle, we moved beyond direct instruction in word knowledge, and included instruction focused on developing one's word-learning ability (Baumann et al., 2002; Baumann & Edwards et al., 2003; Stahl & Fairbanks, 1986; Stahl & Nagy, 2006). Specifically, we focused on building students' morphological awareness, defined as the understanding of complex words as combinations of meaningful smaller units or morphemes (i.e., prefixes, suffixes, and roots) that contribute to the words' meanings and functions (e.g., Kuo & Anderson, 2006). Finally, we drew on the research that suggests that to be effective in promoting one's literacy-related competencies, such instruction must include opportunities for reading, writing, listening and speaking (Beck, McKeown, & Kucan, 2002; Berninger & Abbott, 2010; Snow et al., 1998). By operationalizing these principles we intended to organize the study of language to go deeper than standard practice in the middle school ELA classroom, and as a result, improve literacy outcomes.

Vocabulary Instruction with Linguistically Diverse Adolescents

Historically, research examining vocabulary instruction has largely been conducted with English-only (EO) learners enrolled in primary grade classrooms (for relevant reviews, see National Reading Panel, 2000; Shanahan & Beck, 2006). Yet recent demographic trends, combined with achievement data, highlight the challenge of meeting linguistically diverse adolescents' literacy needs (e.g., Carnegie Council on Advancing Adolescent Literacy, 2010; National Center for Education Statistics, 2012; Short & Fitzsimmons, 2007). In turn, the past decade has seen a relative surge in evaluation studies, reviewed below, focused on vocabulary instruction for LM students and their EO classmates who are beyond the primary grades.

Collectively, these studies focus on students from 5th through 8th grade, investigating the role of online (Dalton, Proctor, Uccelli, & Mo, 2011; Proctor et al., 2011), teacher-delivered (August et al. 2009; Authors, 2010b; Carlo et al., 2004; Lubliner & Smetana, 2005; Snow et al., 2009; Vaughn et al., 2009), and researcher-delivered (Townsend & Collins, 2009) vocabulary instruction, ranging from 5 weeks (Townsend & Collins, 2009) to 24 weeks (Snow et al., 2009) in duration. They also vary with respect to their relationship to standard practice: Seven interventions were implemented as part of the instructional core for a single subject-area (ELA, science, or social studies; August et al.; 2009; Authors, 2010b; Carlo et al., 2004; Dalton et al., 2011; Lubliner & Smetana, 2005; Proctor et al., 2011; Vaughn et al., 2009), one was designed as a school-wide initiative to provide a daily vocabulary lesson as part of each content area class as well as ELA (Snow et al., 2009), and one tested the effects of supplementary vocabulary instruction delivered in an after-school setting (Townsend & Collins, 2009). All nine studies reported significant treatment effects on curriculum-based measures of words taught. Additionally, findings also suggested effects on curriculum-based measures of content taught (August et al., 2009; Vaughn et al., 2009), researcher-developed measures of morphological awareness (Authors, 2010b), metacognitive skills (Lubliner & Smetana, 2005), and reading comprehension (Authors, 2010b; Carlo et al., 2004; Lubliner & Smetana, 2005), as well as significant effects on a norm-referenced measure of reading comprehension (Authors, 2010b). For the seven studies that examined whether language status had a relationship with treatment effects, findings from six suggested that the interventions were equally effective for LM and EO students (August et al. 2009; Authors, 2010b; Carlo et al., 2004; Dalton et al., 2011; Proctor et al., 2011; Vaughn et al., 2009). In contrast, for the school-wide vocabulary teaching initiative designed and evaluated by Snow and colleagues, the LM students benefited more than their EO peers (Snow et al., 2009). Taken together, the findings point to the potential of employing principles of vocabulary instruction—deemed effective in research with young EO students—to classroom settings serving linguistically diverse populations, but further research is needed, particularly that which is large-scale, implemented under typical conditions, and experimental in nature. We note that of these nine recent studies, only three were experimental in nature (August et al., 2009; Townsend & Collins, 2009; Vaughn et al., 2009). Moreover, most of the existing studies focused on determining efficacy based on researcher-developed measures of words taught and reading comprehension, rarely extending beyond these to investigate effects on measures of word-learning, such as morphological skills, or writing, both understudied outcomes worthy of investigation (Graham & Perin, 2007; Pressley, Disney, & Anderson, 2007).

Beyond the Average Treatment Effect: Investigating Student-by-Treatment Interactions

A long line of scholarship examines the theoretical principle that the efficacy of any instructional practice will depend upon the skill level of the individual student (e.g., Cronbach, 1957; Foorman et al., 1998; Tobias, 1976). Indeed, recent research examining student-by-treatment interactions (also referred to as *child characteristics-instruction interactions*; Connor, 2011) has its underpinnings in the theory and methodology of earlier work exploring aptitude-treatment interactions (Cronbach & Snow, 1977). This earlier generation of research produced mixed evidence about such interactions, however, with the benefit of the field's more advanced understanding of developmental processes and the use of increasingly sophisticated analytic strategies to model complex relationships, empirical support for this theoretical principle has accumulated (for a review, see Connor, 2011).

In literacy research, this principle has largely been explored with a lens focused on the ways in which young students' word reading and related decoding skills interact with various approaches to reading instruction (Connor, 2011). This research demonstrates that the skills the student brings to the endeavor influences the degree of impact any given instructional approach has on their reading-skill development (e.g., Connor, Morrison, & Katch, 2004a; Foorman et al., 1998; Juel & Midden-Cupp, 2000). For example, mounting evidence shows that explicit code-based instruction has a greater impact on word reading skills (e.g., phonological awareness, word reading accuracy and fluency) for young students who struggle in these literacy domains than for those who demonstrate above average levels (Connor et al., 2004a; Juel & Midden-Cupp, 2000; Foorman et al., 1998; Sonnenschein, Stapleton, & Benson 2010; Vadasy, & Sanders, 2008).

Much less is known, however, about student-by-treatment interactions in the domains of reading comprehension and writing (Connor, Morrison, & Petrella, 2004b; Olinghouse, 2008; Sonnenschein et al., 2010). While Connor and her colleagues (2004b), for example, found that struggling third grade readers benefited more than their higher-performing peers from teacher-directed explicit instruction, to our knowledge, very little attention has been paid to this principle in evaluation research with adolescents. Klingner and Vaughn (1996) conducted an exploratory study of the effects of a reading comprehension intervention with 26 seventh- and eighth- grade LM students experiencing learning disabilities. They found that students who began the intervention with relatively stronger oral language skills made greater gains on a measure of reading comprehension than their peers who began with lower oral language skills. More recently, in their study to evaluate the effects of vocabulary instruction in an after-school setting (described above), Townsend and Collins (2009) found that of their 37 adolescent LM participants, those with stronger receptive vocabulary skills showed greater response to the intervention, as reflected by their post-test performance on a measure of words taught.

While an overall estimation of a treatment's effect provides an indication of its effectiveness at the population level, it does not shed light on the relative efficacy of the approach. Investigating its effects as a function of variability in students' skills—determining those subgroups for whom the intervention may be particularly effective or for whom it may be less appropriate—is essential to improving the match between instruction and students' needs. Large-scale evaluation research that attends to this principle is needed, particularly that conducted with linguistically diverse adolescents. The intervention under study was designed with an average reader profile in mind, based on prior research with the adolescent reader population in the urban classroom (e.g., Authors, 2010a). While this design reflects an effort to more closely link core instruction to the students' needs, it was, necessarily, still implemented in a diverse context—classrooms characterized by student heterogeneity with respect to vocabulary knowledge and language background in particular. Thus, we used this opportunity to estimate student-by-treatment interactions on these characteristics, in addition to estimating an average effect of the treatment.

Present Study

Designed to build upon and advance research in adolescent literacy, we conducted a randomized field trial to test an academic vocabulary intervention designed to bolster the language and literacy skills of linguistically diverse 6th graders enrolled in 14 urban middle schools. The 20-week intervention was delivered for 45 minutes a day by randomly selected ELA teachers over the course of one academic year. While a series of recent studies, primarily quasi-experimental in nature, suggest vocabulary instruction as a worthwhile endeavor for linguistically diverse students (August et al., 2009; Authors, 2010b; Carlo et al., 2004; Dalton et al., 2011; Lubliner & Smetana, 2005; Proctor et al., 2011; Snow et al., 2009; Townsend & Collins, 2009; Vaughn et al., 2009), the research base in this area remains significantly underdeveloped to inform efforts at scale. Specifically, while it is intuitively appealing to consider that academic vocabulary instruction under typical conditions may be an avenue not only for students' vocabulary development but also for promoting their morphological skills, reading comprehension, and/or written language development, its potential to do so remains unclear. Thus, we designed the present study to address this knowledge gap.

Moreover, even when an impact study of a literacy intervention is conducted, often very little is learned about the extent to which adolescents may have benefited differentially. Yet such knowledge is crucial for the design of targeted intervention, particularly for the struggling reader. Thus, beyond investigating the treatment's overall impact, we also designed the study to investigate the relative efficacy of the intervention—the student performance levels at which response to treatment was greatest—focusing specifically on the role of students' prior vocabulary knowledge and their language background.

Two specific research questions guided this study: 1) What is the impact of an academic vocabulary program on the vocabulary knowledge, morphological skills, reading comprehension, and writing skills of LM students and their EO classmates enrolled in urban middle schools? 2) Does student language background (LM student, EO student) and/or initial vocabulary knowledge differentially predict response to the treatment?

Method

Participants

This cluster-randomized field trial was conducted in 14 middle schools in a large, urban school district in California that serves an economically and linguistically diverse student population. Fifty teachers and 2082 students participated. The cluster-randomized trial meant that each of these 50 teachers was assigned to either the treatment or control condition (the method for random assignment is described below in the *Study Design* section).

Students' home languages were determined based on student surveys administered at pretest. To be sure, language proficiency in one's home language and in English exist on a continuum (Valdés, 2005), but consistent with other research in this area (see below) and for feasibility, students who reported speaking English exclusively at home were classified as EO learners, whereas students reporting that a language other than English was spoken at home to any degree were classified as LM learners. This broad definition for LM learner includes students whose families speak English predominately and those who speak English and another language in equal amounts, as well as those from homes in which another language predominates, consistent with the definition of this population offered by the National Literacy Panel on Language Minority Children and Youth (August & Shanahan, 2006).

Of the students, 71% (*n*=1469) were LM students, 65% of whom (*n*=955) were from homes where Spanish was the primary language. The next most common home languages were Tagalog (11%; *n*=156) and Vietnamese (8%, *n*=114). The proportion of LM students in participating classrooms ranged from 32% to 96%. All participating schools included students living in poverty (median = 51.6%; ranging from 23.0% to 100.0%). Twenty-five teachers along with their students in 39 sections were assigned to the control condition and 25 teachers along with their students in 37 sections were assigned to the treatment condition. There were 971 students in the treatment group (72% LM, *n*=700) and 1111 students in the control group (69% LM, *n*=768).

Missing data. The analytic sample of 2082 students includes all students who were enrolled in the participating classrooms at the time of randomization. Multiple imputation using the Markov Chain Monte Carlo method (Little & Rubin, 1987) was used to account for missing data at pretest or posttest. Specifically, 20 complete datasets were created based on an imputation model that included LM status, ethnicity, gender, and all pretest and posttest scores (with the exception of the writing test which was only administered to a randomly selected subsample of students); twenty datasets rather than the typical 5 to 10 were created because a larger number of datasets is considered to be more appropriate when children are nested in classrooms and schools (Francis, personal communication, 2012). All descriptive and multilevel analyses reported below were conducted with the 20 complete datasets and combined using appropriate procedures to aggregate standard errors. Of the 2082 students enrolled at randomization and included in the analytic sample, 123 (6%) were missing posttest data, largely due to moving out of the participating schools or district. Another 417 students (20%) were missing one or more scores at pretest, largely to logistical difficulties in administering one of the pretests at one school.

It is worth noting that some of this missingness, which involved one missing pretest covariate for an entire school, may be systematic if this school differs in unobserved ways from the other schools in the study. Nonetheless, we decided that including this school would be less likely to bias results than to exclude it, given that it participated in all other aspects of the study. In addition, the robustness of the majority of findings to the use of multiple imputation supports the conclusion that this decision did not seriously bias results.

The results below were all substantively the same when conducted with the multiply imputed datasets or when conducted with the reduced dataset of students with complete data, with the exception of one effect for the writing outcome, as noted below.

Study Design

The cluster-randomized trial required that each of the 50 teachers be assigned to the treatment or control condition. Within each school, teachers were blocked on achievement based on the classroom mean score on the state standards test in ELA administered at the end of the prior academic year, and then randomized within blocks to improve the precision of estimates of treatment effects (Raudenbush, Martinez, & Spybrook, 2007). To conduct random assignment, we employed a random numbers generator to generate, in sequence, random numbers ranging from 0 to 1 that were assigned to the first teacher in each of the blocks. Based on the randomly generated number, the first teacher in each block was assigned to treatment or control and the second teacher in the block was assigned to the other condition.

Since the intervention was designed as an alternative strategy for regular instruction, the amount of time for ELA instruction and the general skills to be taught (i.e., reading, writing, listening, speaking) were comparable across the control and treatment conditions.

Intervention

Academic Language Instruction for All Students (ALIAS) integrates the instructional principles featured in scholarship on vocabulary, reading comprehension, and writing development and instruction. As described, we operationalized these general principles to design a rigorous and engaging approach to academic vocabulary instruction for use in mainstream, low-performing ELA classrooms with high numbers of LM students. In so doing, we drew on reports from individual experimental and quasi-experimental studies and meta-analyses (e.g., August & Shanahan, 2006; Baumann et al., 2002; Baumann & Edwards et al., 2003; Carlo et al., 2004; Graham & Perin, 2007), available research-based books for teachers (e.g, Beck et al., 2002; Graves, 2006; Stahl & Nagy, 2006), as well as promising approaches recommended by practice-based scholarship (e.g., Marzano & Pickering, 2005).

The program was 20 weeks in length, featuring 9 two-week units, each consisting of a 9-day lesson cycle, and 2 one-week review units. Each daily lesson in the cycle was designed to be 45 minutes. These 45 minute lessons were implemented in the context of the participating schools' ELA block, which lasted between 90 and 120 minutes a day. Each unit revolved around a relatively short piece of informational text—a feature article from *Time for Kids* magazine, to which the participating school district subscribes. We selected texts on the basis of several criteria: potential for student engagement, readability at approximately the sixth grade instructional level, length, and the opportunities available for teaching academic vocabulary.

From each text, we chose a set of high-utility academic words, for a total of 70 target vocabulary words. Ten of these words were formally retaught in one or more subsequent units, in the same or different form (e.g., research, researcher). Words that appeared in both a unit text (and were central to comprehending that text) and on the Academic Word List (AWL; Coxhead, 2000) were privileged for study. The AWL is comprised of 570 word families that account for approximately 10% of the total words in academic texts and are outside the first 2,000 most frequently occurring words of English (see the General Service Word List; GSL; West, 1953). This academic domain of vocabulary thus represents high-utility words that appear commonly in expository text, including 6th grade content area textbooks (Nair, 2000) but are not specific to any particular academic discipline. In all, 62 of the words taught appear on the AWL (e.g., expanse, integrated, generate; Coxhead, 2000), 7 words appear on the GSL (e.g., according to, average, social; West, 1953), and 1 word (i.e., inspire) appears on neither list but does appear in two unit texts and exemplifies key characteristics of academic vocabulary.

The first unit, designed to introduce the instructional sequence to the students, focused on 4 words; the other 8 units focused on 8 or 9 academic words throughout the 9 days. To ensure the instruction was authentic, not every target word appears in every day's lesson; rather, selected words were used to process content and meet lesson objectives. The 9-day cycle included a variety of whole-group, small-group, and independent activities designed to incrementally build word knowledge. Each unit begins with exposure to the word in text. Subsequent lessons focus on connections to prior knowledge, additional meanings and uses of the words, morphological analyses, applications of the words in novel contexts, and then, on the final 2 days of the instructional cycle, use of the words in students' own writing. These two writing lessons focused on idea generation and composition and the instructional strategies can largely be considered structured and explicit. The one-week review units included cooperative games focused on the previously taught words as well as opportunities to re-teach specific words.

The program also provided teachers with additional, supportive teaching materials. These included a map of the 9-day cycle, an activity reference sheet, a one-page outline for each daily lesson, and a more elaborated "instructional model"—or sample script—meant to offer further clarity and depth regarding each lesson's content and potential challenges (for more information on the intervention see Authors, 2010b, 2010c, 2010d).

Professional Development

To support the implementation of the intervention, teachers in the treatment group engaged in monthly meetings with one of two program specialists, both former teachers with extensive experience in the district and trained by the research team. The meetings were designed to support teachers' program implementation, and as such, were guided by their specific professional needs. Program specialists addressed questions and issues, from minor logistical details to more substantive challenges, such as troubleshooting difficult lessons. In addition, each teacher in the treatment group had access to a password-protected website, which included electronic versions of the instructional materials and video clips that featured the implementation of each lesson in similar classrooms.

Measures

Student language and reading skills. We administered a battery of standardized and researcher-created assessments in the domains of vocabulary, morphological awareness, reading comprehension and writing to students at pretest (October 2008) and posttest (May 2009). At pre-test, the battery included three standardized measures and two researcher-developed measures, and at post-test the battery included two standardized measures and six researcher-developed measures. For researcher-created assessments, we used forms with the same test items (rearranged in different orders to minimize their surface similarity); though this may introduce practice effects, these effects can be assumed to be the same across treatment and control conditions. The researcher-created assessments were developed and used in prior studies conducted in schools and classrooms with similar demographics and achievement levels to those of the participating schools.

Vocabulary. Vocabulary was assessed using one standardized, norm-referenced measure and three researcher-created instruments.

Gates MacGinitie Reading Vocabulary Test (Form S; MacGinitie, MacGinitie, Maria, & Dreyer, 2000). This standardized, norm-referenced test is a widely used assessment of grade-level reading vocabulary knowledge. For this task, students are asked to identify the synonym for a given word used in a brief sentence. The publisher provides evidence of adequate reliability

(Kuder-Richardson Formula 20 reliability = .86; alternate forms *r* = .77) and extensive evidence of validity.

Academic Word Mastery. This researcher-created vocabulary measure is a 30-item multiple-choice task in which students choose a synonym for a given word drawn from the words taught in the curriculum. A prior study using a version of this task with a similar population provided convergent and divergent validity evidence (Authors, 2010b) and extensive research provides evidence for the validity of this commonly used paradigm for assessing vocabulary (for a review, see Pearson, Hiebert, & Kamil, 2007). The estimate of internal consistency reliability at posttest was high (Cronbach's alpha =.85).

Word Association. Depth of vocabulary knowledge was assessed with the Word Association task. Drawing on the design of tasks used in prior research (Carlo et al, 2004; Schoonen & Verhallen, 1998 as cited in Carlo et al, 2004), this task taps students' knowledge of paradigmatic associations (e.g., that *effect* can be substituted with *consequence* while preserving meaning) and syntagmatic associations (e.g., that an *effect* can be *caused*). The task consisted of 15 items, each of which drew on knowledge of words taught in the program. Each target word appeared in the center of a box, surrounded by 6 other words, 3 of which were immutably associated with the target word, and three of which were only circumstantially related to the target word. For instance, *effect* has immutable associations with *cause*, *consequence*, and *result*, but only circumstantial associations with *negative*, *policy*, and *people*. Students were instructed to choose the three words that "always go with the target word" and provided with feedback on two practice items using common words (i.e., *foot* and *dog*). Students earned a point for each correct association for a possible total score of 45. Consistent with previous research using this measure, no points were deducted for incorrect answers (Authors, 2010b; Carlo et al, 2004). The estimate of internal consistency for this task in a prior study (Authors, 2010b) was adequate (Cronbach's alpha= .78). Prior studies, including Schoonen and Verhallen (1998) and two with similar populations (Authors, 2010b; Carlo et al., 2004), provide convergent and divergent validity evidence.

Academic Word Meanings-in-Context. To complement the more decontextualized vocabulary measures, this task assessed students' comprehension of academic word meanings in the context of extended expository texts. This measure, used in previous research (Authors, 2010b), draws on the framework proposed by Pearson, Hiebert, and Kamil (2007). Specifically, it includes five expository passages drawn from *Time for Kids* that were candidates for inclusion in the program, but did not appear in the final version. As originally written, each passage contained three academic words that had been taught in the instructional program. Students read each passage independently and answered 6 multiple choice questions following each passage, 3 of which were vocabulary questions that tapped understanding of a taught academic word in the context of the passage (e.g., *What does* <u>*major*</u> *mean in the text?* for a passage describing a teenager who can discuss *every* <u>*major*</u> *Presidential candidate from 1896 to 2004*). For these vocabulary items, the correct choice was an appropriate synonym (e.g., *important* for *major,* in this context), whereas the distractors included a different meaning for the target word than that used in the passage (e.g., *military officer*), a word that is related to the content of the passage but is not a meaning of the target word (e.g., *Republican*), and a word that looks similar to the target word and has a loose semantic relationship to the content of the passage (e.g., *majority*). As such, it can be considered a contextualized vocabulary measure with a strong reading comprehension component. The internal consistency reliability for this 15-item task was adequate (Cronbach's alpha = .70). The other 3 questions for each passage focused on reading comprehension skills, and are described below in the section on reading comprehension.

Morphological awareness. Students' morphological skills were assessed using two researcher-created instruments.

Morphological Decomposition. This assessment of morphological awareness was created based on previous research (Carlisle, 2000; Carlo et al., 2004). In this task, testers provide students with a word with a derivational suffix (e.g., complexity) and ask them to extract the base word (e.g., complex) to complete a sentence (e.g., The problem is ___________). The task included 18 items, 9 of which included words taught in the program. When selecting these nine target words, we chose words not included on the Academic Word-Meanings-in-Context measure that could be derived clearly with the three suffixes chosen (*-ity*, *-sion/-tion*, *-al*). When selecting these three suffixes, we chose those taught in the program thought to be the most challenging and useful. To minimize the influence of word-reading skills, test administrators first read the word and sentence aloud, and students responded with a written answer. Trained research assistants scored written answers to the task dichotomously using a detailed scoring guide that included a rubric along with sample correct and incorrect responses. Responses were scored as correct if they provided the correctly spelled form of the word or a phonetically justifiable version of the word form, such as *posess* for *possess* or *durible* for *durable*.

Responses were scored as incorrect if they were morphologically unrelated words such as *have* for *possess* or *hard* for *durable*, when they were incorrectly decomposed responses such as *poss* or *dura*, or when they were ambiguous responses such as *possese* and *durabil*. In this way, we protected, in part, against the confounding of variation in students' ability to spell the base word with true variation in morphological awareness. The estimate of internal consistency reliability at posttest was high (Cronbach's alpha = .84). Several prior studies provide validity evidence for tasks using this paradigm, including convergent and divergent validity evidence (Authors, 2010b; Carlisle, 2000; Carlo et al., 2004) and evidence of construct validity based on confirmatory factor analysis models in both LM and native English-speaking populations (Authors, 2012).

Morphological Derivation. This assessment of morphological awareness was developed based on previous research (Nagy, Berninger, & Abbott, 2006; Tyler & Nagy, 1989). In this task, students complete a sentence (e.g., *The man is a great* _______.) by choosing a nonsense word with an appropriate derivational suffix (e.g., *tranter*) from among four choices (e.g., *tranter*, *tranting*, *trantious*, *trantiful*). The task, sometimes referred to as a suffix choice task, included 18 items. The estimated internal consistency reliability at posttest was adequate (Cronbach's alpha = .77). Several prior studies provide evidence for the validity of tasks using this paradigm, including evidence of convergent and divergent validity in native English-speaking (Nagy, Berninger, & Abbot, 2006) and LM populations (Authors, 2010b) as well as evidence of construct validity based on the fitting of confirmatory factor analytic models in both LM and native English-speaking populations (Authors, 2012).

Reading Comprehension. Reading comprehension was assessed using one researcher-created instrument and one standardized, norm-referenced measure.

Comprehension of Expository Text including Academic Words. This assessment measures global comprehension of expository texts that included academic words taught. This task included five expository passages drawn from *Time for Kids* that were candidates for inclusion in the program, but do not appear in the final version. Students read each passage independently and answered multiple-choice comprehension questions during a 45-minute period. The items for each passage included 3 multiple-choice questions. The first question measured global comprehension of the passage (e.g., *What is the main idea of this text?*), while the second and third item required students to make an inference across several statements in the passage (e.g., After students had read several sentences that describes three people named Alex, Joshua, and Nathan, they were asked *What do Alex, Joshua, and Nathan have in common?*). The estimate of internal consistency reliability for the resulting 15-item task at posttest was adequate (Cronbach's alpha = .74). Prior studies provide evidence for the validity of this task with linguistically diverse populations, including evidence of convergent and divergent validity (Authors, 2010b) and evidence of construct validity based on the fitting of confirmatory factor analytic models (Authors, in press).

Gates-MacGinitie Reading Test, Fourth Edition: Reading Comprehension (6th grade version; MacGinitie et al., 2000). This standardized, norm-referenced measure is a widely used assessment of students' global reading comprehension, in which students have 35 minutes to read several grade-level passages from expository and narrative texts and complete multiple-choice questions. Form S was used at pretest and form T was used at posttest, and equated extended scaled scores were used in analyses. The publisher reports Kuder-Richardson Formula 20 reliability coefficients of .90 to .92 for the sixth grade test, as well as extensive validity evidence.

Writing. Writing was assessed using one standardized, norm-referenced measure.

Oral and Written Language Scales: Written Expression. (Carrow-Woolfolk, 1996). This standardized, norm-referenced measure is an assessment of students' written language skills, in which students respond to structured and open-ended writing tasks that require them to demonstrate their ability to use written English conventions, linguistic forms, and to communicate meaningfully. The examiner reads aloud a verbal stimulus for each of the twelve items, and following each set of directions, students respond in written form. Due to practical constraints, this task was only administered and scored for a sub-sample of students randomly selected from within each participating classroom (*n* = 746; *n* Treatment = 357; *n* Control = 389).

Trained research assistants scored student responses using the detailed guide provided in the test manual. Each response was scored by applying one or more scoring rules related to a specific aspect of the test item. These scoring rules are based on the three writing skill areas assessed (i.e., conventions, linguistics, and content). Item score totals ranged from 2-11 points and students could earn anywhere between 0 and the maximum score for each item. The total possible raw score for the 12-item measure is 70. For the written expression scale, publishers report split-half reliability coefficients of .88 for children age 12 to 13 years.

Fidelity of Implementation. Fidelity of implementation was estimated and assessed using three methods: a weekly log completed by every teacher in the treatment group, classroom observations conducted in the treatment classrooms prior to and during the implementation of ALIAS, and classroom observations conducted in the control classrooms for the study duration.

Implementation logs. Treatment teachers completed implementation logs for each unit-cycle (*n*=11 per teacher, total=275 logs). On each log, teachers recorded the dates spent on the unit and completed nine sections, one for each of the nine daily lessons. Each daily lesson section contained the lesson's objective as well as two sub-sections with the following headers: 1) *approximate minutes spent on ALIAS*, and 2) *activities completed* with a checklist of the two to six lesson components (e.g., *preview and read aloud article, introduce target words*). On each log, there was also space for teachers to provide comments as well as notes on differences between sections. Complete data was available for 99% of the logs.

Classroom observations in treatment classrooms. Trained research assistants observed treatment teachers once prior to the beginning of the intervention, and on five occasions during the intervention period. The observations were supported by a protocol with which they rated teachers' implementation of the intervention for fidelity (i.e., the presence of each of the two to six lesson components for a given daily lesson) and quality (i.e., ratings of low, medium, or high on a range of nine instructional quality variables specific to the intervention). Observations occurred every two to three weeks during the implementation period, scheduled in such a way that, for all teachers, we would capture a range of daily lessons across the lesson cycle and the unit. Reliability was first established during training with video examples, and then 20% of observations in treatment classrooms were conducted in pairs (i.e., double-coded) to estimate reliability observations conducted. Estimates from these observations indicated high inter-rater reliability for fidelity of implementation (percent agreement = 100%; Cohen's Kappa = 1.00) and for the indicators of instructional quality (average percent agreement = 90.3%; Cohen's Kappa = .80). Ratings for fidelity and quality were averaged across items and across observations to yield composite scores for each teacher.

Classroom observations in control classrooms. Control teachers were observed once prior to the start of the intervention period and five times during the intervention period. These observations lasted for 45 minutes and were divided into three 15-minute intervals, with support from an observation protocol featuring 11 categories for curricular content developed based on the categories used in the California state curricular content standards for English-language arts. Observers also coded any vocabulary instruction on the observation protocol, using a coding scheme adapted from previous research (Gersten, Dimino, & Jayanthi, 2007).

After the completion of the 45 minute observation, observers rated the control teachers on 15 instructional quality indicators, equivalent to those used in the treatment classrooms. Of these, 9 indicators were specific to the instructional approach used in intervention, although potentially observable in any classroom (e.g., Teacher affirms correct word definitions/usages; Teacher facilitates student talk) and 6 indicators were related to general instructional quality (e.g., Teacher is prepared for class; Teacher responds effectively to misbehavior). Instructional quality ratings were averaged across items and observations, yielding two composites for each teacher. Inter-rater reliability estimates based on double-coding of 20% of the observations were high for the content code (percent agreement = 100%; Cohen's Kappa = 1.00) and for the indicators of instructional quality (average percent agreement = 95.0%; Cohen's Kappa = .90).

Data Analyses

Multilevel modeling (a.k.a. Hierarchical Linear Modeling) was used to estimate the treatment effect while accounting for the nesting of students within teachers (Raudenbush, 1997; Raudenbush & Bryk, 2002). To evaluate whether the intervention had an impact on students' performance, we fitted a sequence of multilevel models in which the posttest score for each measure was regressed on a dummy variable representing condition (treatment or control) and pretest covariates. Specifically, we fitted two-level models to account for the nesting of children within teachers. These models carry a benefit over simple ANCOVA models in that they produce standard errors and corresponding inference tests that are not biased by dependence among the residuals given the hierarchical organization of the data. To improve the precision of the estimate of the treatment effects, we included the pretest score for the most relevant measure for each outcome at the student level (centered at her or his teacher's mean) and at the teacher level (i.e., teacher mean centered at the grand mean).

Although there were other levels of nesting in the data, such that sections (i.e., class periods) were nested within teacher and teachers were nested within schools, preliminary fitting of the baseline models indicated that the majority of higher-order variation (i.e., variation beyond that at the student level) was at the teacher level across outcomes, rather than at the section- or school-levels. Specifically, results of fitting four-level models (i.e., accounting for school, teacher, section, and student levels), which included the pretest covariates but not the treatment effect, indicated that variation was not significant at the school level for six of the eight outcomes; for the other two outcomes, the four-level model encountered convergence problems suggesting model misfit, perhaps due to the small number of schools. Results of fitting unconditional three-level models (i.e., accounting for teacher, section, and student levels) indicated that across all eight outcomes, variation at the teacher level was substantially larger than variation at the section level; the teacher-level intraclass correlations ranged from .12 to .21, with an average across outcomes of .17, while the section-level intraclass correlation ranged from .01 to .05, with an average across outcomes of .03. Including the effect of treatment in these three-level models, along with pretest covariates, encountered convergence problems for several outcomes, perhaps due to the small number of sections nested within each teacher. Given this evidence that nesting within teacher is the most important level and given that teacher was the level of assignment, we thus report results of the two-level model accounting for nesting of students at the teacher level for all outcomes. However, it is worth noting that all substantive results were the same when alternate two-level models accounting for section rather than teacher were used.

To illustrate this approach, the hypothesized multilevel model for the academic word mastery outcome is given by the following equations. The level-1 (student-level) equation is:

$$POST\_AWM_{ij} = \beta_{0j} + \beta_2 PRE\_AWM_{ij} + (\varepsilon_{ij})$$

$POST\_AWM_{ij}$ represents the posttest score on academic word mastery for child i with teacher j. Term β_{0j} represents the teacher-level intercept for teacher j, which is determined by the level-2 equation in the following paragraph. Term β_2 represents the effect of the student-level pretest (i.e., pretest academic word mastery in this case) for child i with teacher j. Residual ε_{ij} represents the random effect for child i for teacher j, which is drawn from a normal distribution with unknown variance σ_{ε}^2.

The level-2 (teacher-level) equation is:

$$\beta_{0j} = \gamma_{00} + \gamma_1 TREATMENT_j + \gamma_3 AVG\_PRE\_AWM_j + (u_j)$$

Parameter γ_{00} is the overall intercept. Parameter γ_1 represents the main effect of treatment on the posttest score. Term γ_3 represents the effect of the teacher-level average for the pretest. Residual u_j represents the random error in the random intercept for teacher j, which is drawn (independently from ε_{ij}) from a normal distribution with mean 0 and unknown variance σ_1^2. Thus, the intercept was allowed to vary by teacher, while all other effects were specified to be fixed across teachers. In particular, the effect of treatment was fixed to be the same across teachers, consistent with our interest in the overall average effect of the treatment. In addition, the model assumes that the effect of the pre-test is constant across all teachers and across conditions, no heterogeneity of regression, and no random effect of the pre-test.

In a second set of multilevel models, we investigated whether the treatment effect varied as a function of students' language group (i.e., LM or EO) and/or as a function of students' pretest levels of vocabulary (using a standardized measure). For the latter question, we took two approaches. First, to allow for nonlinear interactions and to determine the approximate performance levels at which response to treatment was greatest, the sample was divided into four groups based on whether students' vocabulary scores at the outset of the study were in the first, second, third, or fourth quartile. Second, we tested for linear interactions using the continuous pretest score; this approach has the benefit of parsimony and indicates whether higher levels of pretest vocabulary are associated with greater treatment effects.

To investigate whether the effect of treatment differed for LM students compared to EO students, a dichotomous variable for language group (coded with 1 = EO and 0 = LM) was included in the equation above along with a term for the interaction between treatment and language group. A likelihood ratio test was then conducted to determine if this model fit significantly better than a model with only the main effect of treatment and the main effect of language group. To investigate whether the effect of treatment differed as a function of pretest vocabulary quartile, three dummy variables for quartiles 2, 3, and 4 (with quartile 1, the largest group, serving as the reference category) were included in the equation above along with terms for the interaction between each of these dummy variables and treatment. A likelihood ratio test was then conducted to determine if this model fit significantly better than a model which included the main effects of vocabulary quartile and the main effect of treatment, but not the three interaction terms. For each outcome, we also used a likelihood ratio test to test a linear interaction between treatment and pretest vocabulary. To provide further insight into the differences of effects across sub-groups, additional models were fitted by removing the main effect of treatment and replacing it with the full array of sub-group-by-treatment interaction terms, allowing us to estimate the magnitude and standard error for the treatment effect for each sub-group and thereby to determine if the treatment effect was significant for each sub-group.

Results

Preliminary Analyses: Fidelity of Implementation and Experimental Contrast

There was high fidelity of implementation in the treatment classrooms; on average, participating teachers reported completing 94% (SD=6%; range=22%) of lessons, while observers rated 86.7% (SD=12.5%; range= 45.8%) of lesson components completed. On average, teachers reported teaching ALIAS 47 minutes per day, for a total of 94 days; this is consistent with the design of the program, which included 91 daily lessons to be taught in 45-minute periods over 20 weeks.

Analyses conducted on instructional quality prior to estimating treatment impacts showed that the control teachers were rated as moderately higher than the treatment teachers in general instructional quality ($d = 0.55$), suggesting the evaluation represents a conservative test of the treatment. At the same time, control and treatment teachers were well-differentiated on program-specific aspects of instructional quality; the treatment teachers were rated as much higher (scoring at the medium to high level) than the control teachers on the 9 indicators of instructional quality that were specific to the instructional approach ($d = 1.85$).

Preliminary Results: Student Descriptives

As shown in Table 1, prior to fitting the models we calculated the descriptive statistics for the full sample, to shed light on the language and literacy skills in the population studied and to establish the comparability of the treatment and control groups. We also described student performance by language group (LM students, EO students).

We examined the posttest variables for evidence of ceiling and floor effects, overall and by language group, prior to multiple imputation. In particular, there is reason to be concerned with ceiling effects in posttests, because these can lead to the underestimation of treatment effects or distort treatment-by group interactions (e.g., if the ceiling effects are more pronounced for EO students in this study). For the overall sample, we found some evidence for moderate skewness for Academic Word Mastery (-0.919), for high skewness for Word Association (-1.212), high skewness for Morphological Decomposition (-1.224), and moderate skewness for Morphological Derivation (-0.552). Comprehension of Expository Text including Academic Words, Gates MacGinitie Reading Comprehension, and Written Expression had distributions that were approximately symmetric. We also found some evidence that these ceiling effects were more pronounced for EO than for LM students, with more highly negative skewness on Academic Word Mastery (EO: -1.504; LM: -0.774), Word Association (EO: -1.851; LM: -1.002), Academic Word Meanings-in-Context (EO: -0.502; LM: -0.273), Morphological Decomposition (EO: -1.499; -1.111), Morphological Derivation (EO: -0.782; -0.470), and Comprehension of Expository Text including Academic Words (EO: -0.541; LM: -0.240). Gates-MacGinitie Reading Comprehension and Written Expression were approximately symmetric for both EO and LM learners. The more pronounced ceiling effects evident for EO learners suggest that interactions between treatment and language group should be interpreted with some caution.

Research Question #1: Overall Treatment Impacts

We found significant and meaningfully-sized intervention effects on students' academic vocabulary knowledge, as well as comprehension of expository texts that included academic words and writing. Overall, the effects were greater on word-level measures of vocabulary than on those involving the comprehension of text that included academic words, a more complex assessment. Specifically, as shown in Table 2, the main effect of treatment was significant for a measure of academic word mastery ($d = 0.41$; $p < .0001$); word associations, a measure of depth of word knowledge ($d = 0.22$; $p < .0001$); a measure of academic words presented in text ($d = 0.17$; $p = .0020$); and both measures of word-learning skills, morphological decomposition ($d = 0.40$; $p <.0001$) and morphological derivation ($d = 0.21$; $p < .0001$). The main effect of treatment was significant for a measure of comprehension of expository text including academic words (*d* = .15; *p* = .0076), but not significant for the standardized measures of reading comprehension (*d* = 0.04; *p* = .4256). The main effect of treatment was also significant for the standardized measure of written expression (*d* = 0.19; *p* = .0388). It is worth noting that this last effect had a similar magnitude but was not statistically significant in the reduced sample of participants with complete data, likely due to lower statistical power in the reduced sample.

Research Question #2: Student-by-Treatment Interactions

As shown in Table 3, which present the results of the multi-level modeling to investigate student-by-treatment interactions, our findings suggest that, indeed, the effect of the intervention varied in significant and meaningful ways for certain groups of students. For the measure of academic word mastery, there was a significant interaction between treatment and language group (*p* < .0001), such that the treatment effect for LM learners (*d* = 0.49) was substantially larger than the treatment effect for EO learners (*d* = 0.21), although the treatment effects for both groups were statistically significant.

In addition, for three outcomes, there were significant interactions between treatment and pretest vocabulary levels. Specifically, for one measure of word-learning ability (morphological decomposition), the treatment effects differed significantly by pretest vocabulary quartile (*p* = .0163), such that they were generally larger for students at the lower end of the distribution; effects were substantial and significant for students in the first (*d* = 0.52), second (*d* = 0.33), and third quartile (*d* = 0.27), but did not reach significance for students in the fourth quartile (*d* = 0.24; *p* = .0604). The linear pretest vocabulary-by-treatment interaction term was also significant for this measure of morphological awareness (*p* = .0024), such that higher levels of pretest vocabulary were associated with smaller treatment effects. On the standardized measure of writing, significant treatment effects were found for the first quartile (*d* = .30; *p* = .0059), but not for the other three quartiles. Although the pretest vocabulary-quartile-by-treatment interaction was significant (*p* < .0001), the linear pretest vocabulary-by-treatment interaction term did not reach significance (*p* = .0644), consistent with the nonlinearity of this interaction. For the measure of academic word mastery, both the vocabulary-quartile-by-treatment (*p* = .0274) and the linear vocabulary-by-treatment interaction (*p* = .0113) were significant, such that students with lower pretest vocabulary levels benefitted more than students with higher pretest vocabulary levels. However, when the above-mentioned interaction between language group and treatment was included along with this interaction in a single model, language-group-by-treatment interaction remained significant (*p* =.0006) while the linear vocabulary-by-treatment interaction became non-significant (*p* = .1322). Similarly, when included together, the language-group-by-treatment interaction was significant (*p* = .0002), but the vocabulary-quartile-by-treatment interaction was not (*p* = .1839). These results indicate that language group is a superior predictor of response to treatment for this measure of academic word mastery.

For the measure of text comprehension including academic words, neither the linear vocabulary-by-treatment interaction (*p* = .1681) nor the vocabulary-quartile-by-treatment interaction (*p* = .1073) was significant. However, inspection of magnitude and significance of the effect sizes for the different vocabulary quartile suggests a trend toward differential effects. Specifically, the treatment effects were largest and significant for students in the third quartile (*d* = .29), but were smaller and non-significant for students in the first quartile (*d* = .09), second quartile (*d* = .12) and fourth quartile (*d* = .10). Although these differences in significance may be in part due to differences in sample sizes among the subgroups, the differences in magnitudes suggest that this is a trend worthy of further study.

Discussion

The primary goal of this field-randomized trial was to generate new insights about vocabulary instruction in the context of urban middle school literacy reform. Specifically, the research reported here tested whether an academic vocabulary intervention, designed for use in middle school classrooms with high proportions of language minority (LM) students, would cause stronger language, reading, and writing skills. We interpret the findings from this study as suggesting that one promising avenue for literacy reform in this context is to provide early adolescents with explicit instruction to build up their knowledge of the words they will inevitably encounter in text. Overall, the findings indicate that students benefited from the intervention, particularly those most at-risk. At the same time, the results reinforce and highlight the complexity of improving adolescent literacy rates.

In response to our first research question, we found significant effects on students' academic vocabulary knowledge, morphological awareness skills, written language skills, and comprehension of expository texts that included academic words taught in the program. These overall effects show promise for implementing interventions intended to meet the common language-learning needs of a given student population. Importantly, and consistent with other research in this area, the magnitude of the effect sizes, in part, reflects the skills under focus in the intervention. They were largest for those measures tapping skills "closest" to the intervention itself —also termed near transfer—and diminished in size the "farther" the skills measured got from the intervention's curricular focus (far-transfer). For example, overall, the effects were greater on the word-level measures of vocabulary than on the measure of reading comprehension of expository texts that included academic words.

Given the limited research in this area (Baker et al., 2013; Goldman, 2011), particularly large-scale experimental research conducted under typical conditions (e.g., teacher-delivered, whole-group intervention with minimal additional resources), it remains difficult to interpret the magnitude of the effect sizes with respect to what might be considered a successful intervention (e.g., Bloom, Hill, Black, & Lipsey, 2008). On the one hand, when compared to findings from smaller-scale, highly controlled literacy-related intervention studies conducted with adolescents, the effects demonstrated here are modest (for a review, see Edmonds et al., 2009); however, large-scale, experimental literacy intervention research conducted with adolescents, may prove a better metric for judging the relative practical importance of our effect sizes. To that end, two such school-based studies conducted with large samples report findings that largely mirror the patterns in effect sizes reported here (Kim et al., 2011; Vaughn et al., 2011).

Specifically, these two intervention studies examined the effects of instructional approaches targeting reading (Vaughn et al., 2011) and writing (Kim et al., 2011), but were guided by several shared design principles—principles that also guided the current intervention design. That is, in both cases, the interventions were text-based and anchored learning in rich content; they focused on building students' metacognition, using strategies that oriented students to the *process* of comprehending or generating complex texts; they emphasized interaction among students, providing them with regular opportunities to work and talk together; and both interventions were organized around an instructional routine, providing students with regular opportunities to revisit and practice, and ultimately internalize, strategies over time. Together, the two interventions tested produced moderate, significant effects on the literacy domains under study (Kim et al., 2011; Vaughn et al., 2011). As noted, the present study tested an intervention guided by similar design principles, but focused on vocabulary as the domain under study. In this way, we build off of and extend the findings from this existing evidence, and suggest that a consistent text-based, process-oriented, and interactive approach to learning also benefits adolescents' vocabulary development. Taken together, a question for future research is whether an even more comprehensive approach, guided by these principles and but focusing on reading, writing, and oral language, would further enhance adolescents' literacy development.

The significant effect of treatment on the standardized measure of written expression, while small in magnitude, is of particular note in light of the need for research to investigate the impact of vocabulary instruction on writing outcomes (Graham & Perin, 2007; Pressley et al., 2007). There are several ways to interpret this somewhat surprising finding. First, this finding might be attributable to the particular literacy competencies under focus in the intervention. As previously described, each instructional unit culminated in 2 days focused on developing a written text. Given large-scale research that demonstrates a relative dearth of time spent on writing instruction (Gilbert & Graham, 2010), it may be that the comparably substantial time spent engaged in a structured and explicit writing routine, combined with an increased emphasis on productive vocabulary knowledge, yielded these particular gains. Second, we also highlight that the program had significant effects on morphological awareness skills—our largest effect size after that associated with words taught. There is some evidence to suggest that developing this facet of metalinguistic awareness is related to improved written language performance (Carlisle, 1996; Nagy et al., 2006). Given that morphological awareness has been identified as an area of weakness for LM students in particular (Authors, 2012, in press), it may be that this competency was a source of written language weakness that, when strengthened, contributed to gains in students' written language performance.

A long line of scholarship documents the theoretical principle that the efficacy of any instructional practice will depend upon the skill level of the individual student; however, as mentioned, very little attention has been paid to this principle in evaluation research. Yet in addressing our second research question, we found that, indeed, the effects of the intervention varied in significant ways for certain groups of students, suggesting that a student's language background and initial vocabulary knowledge are likely to predict the impact of a language-based academic intervention.

Specifically, our results suggest that comprehensive, ongoing instruction in word knowledge and word-learning strategies are particularly important for adolescent LM students and their EO classroom peers who demonstrate under-developed vocabulary knowledge. For these at-risk learners, the intervention's effects were significant and stronger on near transfer measures (e.g., academic vocabulary knowledge and morphological awareness skills), suggesting that the intervention's design was a strong match. In addition, our findings indicate that this sub-group of students showed gains on the measure of written language. We interpret this interaction as suggestive that the overall treatment impact on writing was likely driven by the magnitude of the gains demonstrated by those students most at-risk. In light of previous research, albeit with younger children, it is not surprising that these at-risk students benefited most from the intervention's explicit, targeted instruction (e.g., Connor et al., 2004b; Olinghouse, 2008).

Importantly, those students at lower risk—students in the third quartile of vocabulary knowledge —also appeared to benefit from the intervention in distinct ways. Specifically, on the measure of comprehension of texts containing academic words, there was a trend toward differential effects in favor of this group; an earlier exploratory study with a comparable population (Klingner & Vaughn, 1996) indeed demonstrated this finding. We also highlight that, for this group, the program also had a significant effect on the Word Association task, our measure of depth of vocabulary knowledge. In turn, we hypothesize that those learners with average vocabulary levels benefited from the opportunity to build a deeper understanding of the abstract words under study, perhaps progressing from a narrow and context-bound understanding of these academic words to a more rich and nuanced understanding of the concepts that these words represent. With increased depth of word knowledge in mind, it stands to reason that, in the context of this intervention, in order for vocabulary instruction to increase a reader's ability to make meaning from text, the student needed sufficiently developed vocabulary knowledge at the program's outset upon which to build. Still, we note that this finding contributes to a small, inconclusive body of research regarding student-by-treatment interactions in the domain of reading comprehension; to be sure, more research is needed.

Finally, highlighting both a developmental and methodological issue, we note that observed treatment effects on text-level comprehension were relegated to the researcher-developed measure of comprehension, comprised of expository passages; the treatment did not show effects on the standardized measure of reading comprehension containing both expository and narrative passages, and containing more rare words than those in the expository measure. The distal measure was selected because it is standardized and widely used; however, some research finds it to be less sensitive for detecting intervention effects because of its global nature (RAND Reading Study Group, 2002; Pearson et al., 2007). Indeed, as described above, vocabulary intervention research has historically shown effects on local measures of near transfer, but rarely do these effects translate into significant gains on far transfer and global measures (Elleman et al., 2009). Still, it is crucial for evaluation research to include standardized, norm-referenced assessments. In addition, reports underscore the importance of developing more tailored and sensitive measures of various literacy skills designed for use with diverse populations (e.g., Morsy, Kieffer, & Snow, 2010; Pearson et al., 2007).

It is also clear, however, that the impact estimates for the two reading comprehension measures provide mixed evidence of transfer, leaving the effect of academic vocabulary instruction on reading comprehension unclear. For some, these results might call vocabulary instruction into question as a means of improving reading comprehension. However, an alternative interpretation of our results is that this instructional approach is a step in the right direction for young adolescents whose literacy difficulties stem from underdeveloped vocabulary knowledge, but raising questions about the intensity (e.g., dosage and duration) needed to translate into improved reading comprehension outcomes. It may well be that a program such as this provided for the entire academic year or, ideally, even across multiple years would show stronger effects. This may particularly be the case for the students with the lowest levels of vocabulary, who benefited from the intervention at the level of word knowledge and who may require further vocabulary development before they are likely to benefit from instruction aimed more squarely on reading comprehension. Ultimately, given the magnitude of the problem of middle school literacy reform, and the questions raised here, what appears likely is that developing adolescents' academic vocabulary is a promising avenue for improving literacy rates, but as part of a multi-faceted, rigorous approach.

Limitations and Next Steps

The findings from this study raise several questions and issues that should inform future research. With respect to the overall study design, guided by the goal of studying classroom-based instruction with high ecological validity in a way that can inform efforts at scale, the approach featured here includes bundled instructional elements to build up word knowledge and word-learning strategies. Going forward, there is a need to employ multiple treatment conditions to isolate those instructional elements and strategies that are most effective in advancing students' literacy outcomes. Relatedly, an experimental study examining the effects of instruction that target different types of words is necessary to document whether the nature of the words chosen for study as they relate to the particular text at hand has an impact on literacy outcomes.

With respect to our findings, we note that there was some evidence of ceiling effects for some measures at posttest; these effects may have led to smaller effect sizes than would have been otherwise found, and, there was some evidence that the ceiling effects were more pronounced for EO than for LM participants. Given this, the student language group-by-treatment interaction on the measure of academic words should be interpreted with caution. Future research is needed to replicate the finding that LM students benefit more from academic vocabulary instruction of the type investigated here.

In addition, while we did demonstrate short-term program effects on specific measures of vocabulary and reading comprehension administered immediately after the 20-week intervention, further research is needed to determine whether there are long-term effects of the program. And finally, though the randomized design of the study is one of its strengths, it does not provide insights into the reform process. Indeed, there are questions about the factors that influence the middle-school teacher's buy-in, uptake, and sustained use of this instructional approach. Such questions could be answered by taking a mixed-methods or qualitative approach to the study design (Authors, 2010b).

Finally, when drawing lessons learned to inform next steps for educators, we would be remiss if we did not remind the reader of the nature of the academic vocabulary instruction developed and tested. This approach integrates grade-level reading and writing objectives alongside listening and speaking, and focuses on building up word knowledge in the context of a unit of study. The challenge for the field, then, is to ensure that the instructional problems of word-learning strategies and vocabulary teaching do not take on a life of their own, conducted in isolation, such as the use of decontextualized lists and discrete strategies. The analogous scenario has been observed in the line of research examining comprehension strategy instruction. Namely, recent research has found that strategy-based comprehension teaching is increasingly characterized by covering and practicing isolated skills and generic strategies at the expense of a focus on content-based learning from text (e.g., Dewitz, Jones, & Leahy, 2009). In contrast to current practice, providing rich, research-based vocabulary instruction as part of literacy improvement efforts will be essential to improving the academic outcomes of the underserved students in today's linguistically diverse classrooms.

References

Alexander, P.A., & Jetton, T.L. (2000). Learning from text: A multidimensional and developmental perspective. In R. Barr, M. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research* (Vol. 3, pp. 285–310). Mahwah, NJ: Erlbaum.

August, D., Branum-Martin, L., Cardenas-Hagan, E., & Francis, D.J. (2009). The impact of an instructional intervention on the science and language learning of middle grade English language learners. *Journal of Research on Educational Effectiveness, 2*(4), 345-376.

August, D., & Shanahan, T. (Eds.) (2006), *Developing literacy in second-language learners: Report of the National Literacy Panel on language-minority children and youth.* Mahwah, NJ: Lawrence Erlbaum Associates.

Authors (2010a).

Authors (2010b).

Authors (2010c).

Authors (2010d).

Authors (2011).

Authors (2012).

Authors (in press).

Baker, S., Lesaux, N., Jayanthi, M., Dimino, J., Proctor, P., Morris, J., Gersten, R., Haymond, K., Kieffer, M. J., Linan-Thompson, S., & Newman-Gonchar, R. (2013). *Teaching English learners in the elementary and middle grades*. Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Retrieved from the NCEE website: http://ies.ed.gov/ncee/wwc/publications_reviews.aspx#pubsearch

Baumann, J.F. (2009). The intensity of vocabulary instruction and the effects on reading comprehension, *Topics in Language Disorders, 29*(4), 312-328.

Baumann, J. F., Edwards, E. C., Boland, E. M., Olejnik, S., & Kame'enui, E. J. (2003). Vocabulary tricks: Effects of instruction in morphology on fifth-grade students' ability to derive and infer word meanings. *American Educational Research Journal, 40*(2)*,* 447-494.

Baumann, J., Edwards, E., Font, G., Tereshinski, C., Kame'enui, E., & Olejnik, S. (2002). Teaching morphemic and contextual analysis to fifth-grade students. *Reading Research Quarterly, 37*(2), 150-176.

Baumann, J.F., & Graves, M.F. (2010). What is academic vocabulary? *Journal of Adolescent & Adult Literacy, 54*(1), 4–12.

Baumann, J. F., Kame'enui, E. J., & Ash, G. (2003). Research on vocabulary instruction: Voltaire redux. In J. Flood, D. Lapp, J. R. Squire, & J. Jensen, (Eds.), *Handbook of research on teaching the English Language Arts* (2nd ed., pp. 752–785). Mahwah, NJ: Lawrence Erlbaum.

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

Beers, S.F., Nagy, W.E. (2011). Writing development in four genres from grades three to seven: syntactic complexity and genre differentiation, *Reading and Writing, 24*(2), 183-202.

Berninger, V. W., & Abbott, R. D. (2010). Listening comprehension, oral expression, reading comprehension, and written expression: Related yet unique language systems in grades 1, 3, 5, and 7. *Journal of Educational Psychology*, *102*(3), 635-651.

Best, R. M., Floyd, R. G., & Mcnamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, *29*(2), 137-164.

Bloom, H.S., Hill, C.J., Black, A.R., & Lipsey, M.W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions, *Journal of Research on Educational Effectiveness, 1*(4), 289-328.

Buly, M. R., & Valencia, S. W. (2002). Below the bar: Profiles of students who fail state assessments. *Educational Evaluation and Policy Analysis, 24*(3), 219–239.

Catts H. W., Adlof, S. M., & Weismer S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading*. Journal of Speech, Language, and Hearing Research, 49*(2), 278–293.

Carlisle, J. F. (1996). An exploratory study of morphological errors in children's written stories. *Reading and Writing: An Interdisciplinary Journal*, *8*(1), 61-72.

Carlisle, J.F. (2000). Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and Writing: An Interdisciplinary Journal*, *12*(3), 169–190.

Carlo, M.S., August, D., McLaughlin, B., Snow, C.E., Dressler, C., Lippman, D.N., et al. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly*, *39*(2), 188– 215.

Carnegie Council on Advancing Adolescent Literacy. (2010). *Time to act: An agenda for advancing adolescent literacy for college and career success*. New York, NY: Carnegie Corporation of New York.

Carrow-Woolfolk, E. (1996). *Oral and Written Language Scale.* Minneapolis, MN: Pearson.

Chall, J. S., Jacobs, V, & Baldwin, L. (1990). *The reading crisis: Why poor children fall behind.* Cambridge, MA: Harvard University Press.

Connor, C.M. (2011). Child characteristics-instruction interactions: Implications for students' literacy skills development in the early grades. In S.B. Neuman & D.K. Dickinson (Eds.) *Handbook of early literacy research* (Vol. 3, pp 256-275). New York, NY: Guilford Press.

Connor, C. M., Morrison, F. J., & Katch, E. L. (2004a). Beyond the reading wars: The effect of classroom instruction by child interactions on early reading. *Scientific Studies of Reading*, *8*(4), 305–336.

Connor, C.M., Morrison, F. J., & Petrella, J. N. (2004b). Effective reading comprehension instruction: Examining child by instruction interactions. *Journal of Educational Psychology*, *96*(4), 682–698.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Dalton, B., Proctor, C.P., Uccelli, P., Mo E., & Snow, C.E. (2011). Designing for diversity: The role of reading strategies and interactive vocabulary in a digital reading environment for fifth-grade monolingual English and bilingual students. *Journal of Literacy Research, 43*(1), 68-100.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Cronbach, L.J., Snow, R.E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York, NY: Irvington.

Dewitz, P., Jones, J., & Leahy, S. (2009). Comprehension strategy instruction in core reading programs. *Reading Research Quarterly, 44*(2), 102-126.

Duncan, G. J., & Murnane, R. J. (Eds.). (2011). *Whither opportunity?: Rising inequality, schools, and children's life chances*. New York, NY: Russell Sage Foundation.

Duke, N. K., & Carlisle, J. (2011). The development of comprehension. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.),*Handbook of reading research* (Vol. 4, pp. 199–228). New York, NY: Routledge.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C. K., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). Synthesis of reading interventions and effects on reading outcomes for older struggling readers. *Review of Educational Research*, *79*(1), 262–300.

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 Educational Effectiveness, 2*(1), 1–44.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, 90(1), 37-55.

Gersten, R., Dimino, J., & Jayanthi, M. (2007). Towards the development of a nuanced classroom observational system for studying comprehension and vocabulary instruction. in B. Taylor & J. Ysseldyke (Eds.) *Educational interventions for struggling readers K-6* (pp. 381–425). New York, NY: Teachers College Press.

Gilbert, J., & S., Graham (2010). Teaching writing to elementary students in grades 4-6: A national survey. *Elementary School Journal, 110*(4), 494-518.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students, *Journal of Educational Psychology, 99*(3), 445-476.

Graves, M.F. (2006). *The vocabulary book: Learning and instruction*. New York, NY: Teachers College Press.

Goldenberg, C. (2011). Reading instruction for English language learners. In M. L. Kamil, P. D. Pearson, E. Birr Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 684–710). New York, NY: Routledge

Goldman, S. R. (2012). Adolescent literacy: Learning and understanding content. *The Future of Children*, *22*(2), 89-116.

Hock, M. F., Brasseur, I. F., Deshler, D. D., Catts, H. W., Marquis, J. G., Mark, C. A., & Stribling, J. (2009). What is the reading component skill profile of adolescent struggling readers in urban schools?. *Learning Disability Quarterly*, *32*(1), 21-38.

Juel, C., & Minden-Cupp, C. (2000). Learning to read words: Linguistic units and instructional strategies. *Reading Research Quarterly*, *35*(4), 458–492.

Kim, J., Olson, C. B., Scarcella, R., Kramer, J., Pearson, M., van Dyk, D.,...Land, R. (2011). A randomized experiment of a cognitive strategies approach to text-based analytical writing for mainstreamed Latino English language learners in grades 6 to 12. *Journal of Research on Educational Effectiveness*, *4*(3), 231–263.

Kintsch, W., & Rawson, K.A. (2005). Comprehension. In M.J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 209-226). Oxford, English: Blackwell Publishing.

Kuo, L., & Anderson, R. C. (2006). Morphological awareness and learning to read: A cross-language perspective. *Educational Psychologist, 41*(3), 161-180.

Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology*, *95*(2), 211.

Lubliner, S. & Smetana, L. (2005). The effects of comprehensive vocabulary instruction on Title 1 students' metacognitive word-learning skills and reading comprehension, *Journal of Literacy Research*, *37*(2), 163-199.

MacGinitie, W., MacGinitie, R., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Test* (*4th ed.*). Itasca, IL: Riverside Publishing Company.

Marzano, R.J., & Pickering, D.J. (2005). *Building academic vocabulary: Teacher's manual*. Alexandria, VA: Association for Supervision and Curriculum Development.

McCutchen, D. (2006). Cognitive factors in the development of children's writing. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 115–130). New York, NY: Guilford Press.

McNamara, D.S., Crossley, S.A., & McCarthy, P.M. (2010). Linguistic features of writing quality, *Written Communication, 27*(1), 57-86.

Morsy, L., Kieffer, M., Snow, C.E. (2010). *Measure for measure: A critical consumers' guide to reading comprehension assessments for adolescents.* New York, NY: Carnegie Corporation of New York.

Murnane, R., Sawhill, I., & Snow, C. (2012). Literacy challenges for the twenty-first century: Introducing the issue. *The Future of Children*, *22*(2), 3-15.

Nagy, W.E., Berninger, V.W., & Abbott, R.D. (2006). Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. *Journal of Educational Psychology, 98*(1)*,* 134-147.

Nagy, W. & Scott, J. (2000) Vocabulary processes. In M.L. Kamil, P.B. Mosenthal, P.D. Pearson, & R. Barr. (Eds.), *Handbook of reading research* (Vol. 3, pp. 269-283). Mahwah, NJ: Erlbaum.

Nagy, W.E., Hiebert, E.H. (2010). Toward a theory of word selection. In M.L. Kamil, P.D. Pearson, E.B. Moje, & P.P. Afflerbach (Eds.), *Handbook of reading research* (Vol. 4, pp. 388–404). New York, NY: Longman.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*(1), 91–108.

Nair, M. (2007). *An analysis of the words appearing in middle school textbooks.* Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Nation, K., Cocksey, J., Taylor, J. S., & Bishop, D. V. (2010). A longitudinal investigation of early reading and language skills in children with poor reading comprehension. *Journal of Child Psychology and Psychiatry*, *51*(9), 1031-1039.

National Center for Education Statistics (2012). *The nation's report card: Vocabulary results from the 2009 and 2011 NAEP reading assessments* (NCES 2013 452). Institute of Education Sciences. Washington, DC: United States Department of Education.

National Reading Panel. (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (National Institute of Health Publication No. 00-4769). Washington, DC: National Institute of Child Health and Human Development.

Olinghouse, N. G. (2008). Student- and instruction-level predictors of narrative writing in third-grade students. *Reading and Writing Quarterly, 21*(1)*,* 3–26.

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*(2), 282–296.

Pressley, M., Disney, L., & Anderson, K. (2007). Landmark vocabulary instructional research and the vocabulary instructional research that makes sense now. In R.K. Wagner, A.E. Muse, & K.R. Tannenbaum (Eds.) *Vocabulary acquisition: Implications for reading comprehension* (pp 205-231). New York, NY: Guilford Press.

Proctor, C.P., Dalton, B., Uccelli, P., Biancarosa, G., Mo, E., Snow, C., & Neugebauer, S. (2011). Improving comprehension online: effects of deep vocabulary instruction with bilingual and monolingual fifth graders. *Reading and Writing: An Interdisciplinary Journal, 24*(5), 517-544.

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

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173-185.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1), 5-29.

RAND Reading Study Group. (2002). *Reading for understanding: Toward a R&D program in reading.* Arlington, VA: RAND Corporation.

Saddler, B., & Graham, S. (2007). The relationship between writing knowledge and writing performance among more and less skilled writers. *Reading and Writing Quarterly, 23*(3), 231–247.

Schoonen, R., Verhallen, M. (1998). Kennis van woorden: De toetsing van diepe woordkennis (Knowledge of words: Testing deep word knowledge), *Pedagogische StudiËn*, 75, 153-168.

Scott, J.A., Jamieson-Noel, D., & Asselin, M. (2003). Vocabulary instruction throughout the day in twenty-three Canadian upper-elementary classrooms. *The Elementary School Journal*, *103*(3), 269–283.

Shanahan, T., & Beck, I. L. (2006). Effective literacy teaching for English-language learners. In D. August & T. Shanahan (Eds.), *Developing Literacy in Second-Language Learners: Report of* *the National Literacy Panel on Language-Minority Children and Youth* (pp. 415–488). Mahwah, NJ: Erlbaum.

Short, D. & Fitzsimmons, S. (2007). *Double the work: Challenges and solutions to acquiring language and academic literacy for adolescent English language learners.* New York, NY: Carnegie Corporation.

Sonnenschein, S., Stapleton, L., & Benson, A. (2010). The relation between the type and amount of instruction and growth in children's reading competencies. *American Educational Research Journal, 47*(2), 358–389.

Snow, C.E., Lawrence, J.F., & White, C. (2009). Generating knowledge of academic language among urban middle school students, *Journal of Research on Educational Effectiveness*, *2*(4), 325-344.

Snow, C.E., & Uccelli, P. (2009). The challenge of academic language. In D.R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 112–133). New York, NY: Cambridge University Press.

Stahl, S.A., & Fairbanks, M.M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*(1), 72-110.

Stahl, S.A., & Nagy, W.E. (2006). *Teaching word meanings.* Mahwah, NJ: Erlbaum.

Storch, A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology, 38*(6), 934–947.

Tobias, S. (1976). Achievement treatment interactions. *Review of Educational Research*, *46*(1), 61-74.

Townsend, D. & Collins, P. (2009). Academic vocabulary and middle school English learners: An intervention study. *Reading and Writing*, *22*(9)*,* 993-1019.

Tyler, A., & Nagy, W. (1989). The acquisition of English derivational morphology. *Journal of Memory and Language*, *28*(6), 649–667.

Vadasy, P. F., & Sanders, E. A. (2008). Repeated reading intervention: Outcomes and interactions with readers' skills and classroom instruction. *Journal of Educational Psychology*, 100(2), 272-290.

Valdés, G. (2005). Bilingualism, heritage language learners, and SLA research: Opportunities lost or seized?. *The Modern Language Journal*, *89*(3), 410-426.

Vaughn, S., Klingner, J. K., Swanson, E. A., Boardman, A. G., Roberts, G., Mohammed, S. S., & Stillman-Spisak, S. J. (2011). Efficacy of Collaborative Strategic Reading with middle school students. *American Educational Research Journal, 48*(4), 938–964.

Vaughn, S., Martinez, L.R., Linan-Thompson, S., Reutebuch, C.K., Carlson, C.D., & Francis, D.J. (2009). Enhancing social studies vocabulary and comprehension for seventh-grade English language learners: Findings from two experimental studies. *Journal of Research on Educational Effectiveness, 2*(4), 297-324.

Vellutino F. R., Tunmer W. E., Jaccard J. J., & Chen R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading, 11*(1), 3–32.

West, M. (1953). A general service list of English words. London: Longman, Green.

Table 1

Descriptive Statistics on Language and Literacy Measures for the Overall Sample (OS; Treatment n = 971; Control n = 1111), Language Minority (LM) Learners (Treatment n = 700; Control n = 768), and English Only (EO) learners (Treatment n = 271; Control n = 343), by Treatment and Control

Measure	Pretest						Posttest					
	Treatment Mean (*SD*)			Control Mean (*SD*)			Treatment Mean (*SD*)			Control Mean (*SD*)		
	OS	EO	LM	OS	EO	LM	OS	EO	LM	OS	EO	LM
Academic Word Mastery	19.19	21.14	18.46	19.04	21.78	17.85	23.32	24.18	23.01	21.08	23.65	19.95
(Raw score out of 30)	(5.62)	(5.61)	(5.44)	(5.70)	(5.09)	(5.50)	(5.09)	(5.44)	(4.89)	(5.78)	(4.94)	(5.73)
Word Association							36.57	37.64	36.17	35.35	36.66	34.77
(Raw score out of 45)							(4.90)	(5.06)	(4.76)	(5.71)	(6.11)	5.42)
Academic Word							9.01	9.73	8.75	8.46	9.42	8.05
Meanings-in-Context							(3.15)	(3.16)	(3.11)	(2.99)	(2.76)	(2.98)
(Raw score out of 15)												
Morphological							14.45	15.05	14.23	13.15	14.43	12.61
Decomposition							(3.51)	(3.12)	(3.62)	(4.31)	(3.74)	(4.39)
(Raw score out of 18)												
Morphological Derivation	10.43	11.18	10.15	10.68	12.07	10.09	12.66	13.44	12.37	12.13	13.07	11.75
(Raw score out of 18)	(4.06)	(3.96)	(4.07)	(4.06)	(3.71)	(4.04)	(3.58)	(3.42)	(3.60)	(3.85)	(3.61)	(3.87)
Comprehension of							9.13	9.95	8.84	8.93	9.92	8.50
Expository Text including							(3.41)	(38.70)	(3.35)	(3.20)	(3.05)	(3.17)
Academic Words												
(Raw score out of 15)												
Gates MacGinitie Reading	493.68	503.27	490.12	496.15	508.32	490.92	497.03	504.43	494.31	498.74	508.68	494.46
Comprehension	(32.19)	(34.10)	(30.64)	(31.89)	(33.91)	(29.56)	(33.96)	(38.70)	(31.45)	(32.38)	(35.38)	(29.86)
(Extended Scale Score)												
Written Expression							36.17	38.20	35.52	34.86	38.86	33.31
(Raw score out of 70)							(10.51)	(9.28)	(10.80)	(11.54)	(10.94)	(11.27)
							n = 357	*n* = 86	*n* =	*n* =	*n* =	*n* =
									271	389	114	275

Table 2

Final Fitted Multilevel Models for Average Treatment Effect of the Intervention on Language and Literacy Outcomes (N = 2082)

	Academic Word Mastery	Word Association	Academic Word Meanings-in-Context	Morphological Decomposition	Morphological Derivation	Comprehension of Expository Text including Academic Words	Gates MacGinitie Reading Comprehension	Written Expression[4]
Fixed Effects								
Intercept	21.10***	35.39***	8.48***	13.06***	12.03***	8.79***	497.23***	34.76***
Treatment	2.26***	1.15***	0.53**	1.59***	0.77***	0.49**	1.43	2.15*
Pretest (Teacher Average)	0.86***[1]	0.78***[1]	0.53***[1]	0.97***[2]	0.89***[2]	0.11***[3]	1.09***[3]	0.26***[3]
Pretest (Student-level)	0.65***[1]	0.54***[1]	0.28***[1]	0.49***[2]	0.56***[2]	0.06***[3]	0.64***[3]	0.18***[3]
Random Effects								
Teacher	0.65**	0.31*	0.22**	0.64***	0.16*	0.22**	23.39**	7.68**
Student	13.62***	16.70***	5.33***	9.48***	7.36***	5.84***	510.3***	75.70***

[1]Pretest Target Word Mastery; [2] Pretest Morphological Nonword Derivation; [3] Pretest Gates Reading Comprehension; [4]*n* = 746
*p < .05; **p <.01; ***p <.001

Table 3

Effect Size Estimates for Treatment Effects, On Average and by Language Group and Pretest Vocabulary, with Associated Log-likelihood Ratio Tests for the Statistical Significance of Interactions (N = 2082)

	Academic Word Mastery	**Word Association**	**Academic Word Meanings-in-Context**	**Morphological Decomposition**	**Morphological Derivation**	**Comprehension of Expository Text including Academic Words**	**Gates MacGinitie Reading Comprehension**	**Written Expression[1]**
Overall	0.41***	0.22***	0.17**	0.40***	0.21***	0.15*	0.04	**0.19***
LM	0.49***	0.19***	0.18**	0.43***	0.18***	0.16**	0.04	**0.21***
EO	0.21**	0.26***	0.17*	0.33*	0.27***	0.13	-0.04	**0.04**
Vocabulary Pretest Q1	0.52***	0.30***	0.10	0.52***	0.27***	0.09	0.02	**0.30****
Vocabulary Pretest Q2	0.39***	0.13	0.21**	0.33***	0.18**	0.12	-0.01	**0.13**
Vocabulary Pretest Q3	0.36***	0.20*	0.30***	0.27**	0.13	0.29***	0.10	**-0.17**
Vocabulary Pretest Q4	0.22*	0.20	0.24	0.24	0.11	0.10	0.13	**0.04**
Log-likelihood Ratio Test for Significance of Interactions								
Treatment by LM (*df* = 1)	16.97***	0.98	0.04	1.25	1.47	0.29	1.62	**1.49**
Treatment by Vocabulary Quartile (*df* = 3)	9.15*	4.17	5.27	10.29*	3.72	6.09	2.92	**22.51*****
Treatment by Vocabulary (Continuous) (*df* = 1)	6.42*	1.79	4.38*	9.22**	1.7	1.9	3.42	**3.21**

[1] *n* = 746; *p < .05; **p <.01; ***p <.001