American Educational Research Journal
August 2015, Vol. 52, No. 4, pp. 750–786
DOI: 10.3102/0002831215579485

Word Generation Randomized Trial: Discussion Mediates the Impact of Program Treatment on Academic Word Learning

Joshua F. Lawrence
University of California, Irvine
Amy C. Crosson
University of Pittsburgh
E. Juliana Paré-Blagoev
George Washington University
Catherine E. Snow
Harvard University

Classroom discussion, despite its association with good academic outcomes, is exceedingly rare in U.S. schools. The Word Generation intervention involves the provision of texts and activities to be implemented across content area class, organized around engaging and discussable dilemmas. The program was evaluated with 1,554 middle grade students in 28 schools randomly assigned to treatment or control conditions. There were large effects on classroom discussion quality across all content areas, especially in math and science (Cohen's d *= 0.38-1.13). The program also produced*

Joshua F. Lawrence is assistant professor at the University of California, Irvine, 3200 Education Building, Irvine, CA 92697; e-mail: *jflawren@uci.edu*. His research focuses on understanding and improving literacy development and educational outcomes of students at risk, including language minority students.

Amy C. Crosson is a research associate at the Learning Research and Development Center, University of Pittsburgh. She studies literacy interventions to support comprehension, academic writing, and academically productive classroom talk for linguistically diverse students in underresourced schools.

E. Juliana Paré-Blagoev is senior research scientist at George Washington University and director of the Center for Applied Developmental Science and Neuroeducation in the Special Education and Disabilities Program. She works at the intersection of neuroscience, developmental psychology, and education to bridge research and practice.

Catherine E. Snow is the Patricia Albjerg Graham Professor at the Harvard Graduate School of Education. She studies language and literacy development in children from birth through adolescence, with particular focus on children at educational risk because of socioeconomic or immigrant status, limited English skills, and/or inadequate schooling opportunities.

significant, though small, effects on taught vocabulary (effect size = .25, p < .01) but no effects on a standardized assessment of general vocabulary. Quality of classroom discussion mediated 14% of the treatment effect on vocabulary outcomes.

KEYWORDS: vocabulary, discussion, middle school, mediation, academic language

Review of the Literature

Introduction

In a randomized trial, we assessed a program that introduced target vocabulary words in a brief text, highlighted them by providing student-friendly definitions, and offered structured, interactive classroom activities in which the words could authentically be used by teachers and students. We evaluated the impact of this program on student knowledge of taught words, student general vocabulary knowledge, and the quality of classroom discussion. We also conducted a mediation analysis to determine if improved discussion was a mechanism accounting for program impacts on vocabulary. We believe that the discussion-based classroom activities are an innovative aspect of the program and a pedagogical approach not, to our knowledge, previously implicated in vocabulary learning.

The curricular materials provided by the program focus discussion on civic and moral dilemmas (not, it is important to note, on the words to be learned). Discussion-based activities were embedded into a program designed primarily to teach vocabulary with a number of pedagogical goals: to provide engaging activities, motivate students to read texts containing the target words in rich semantic contexts, and ensure opportunities for the students to use the newly learned words. These features all characterize successful vocabulary instruction (e.g., Beck, McKeown, & Kucan, 2013; Graves, 2000).

Our exploration of discussion as an instructional feature promoting vocabulary learning links to the large and growing interest in the value of classroom discussion as a factor promoting students' academic skills in a number of different domains. A meta-analysis (Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009) documented the contributions of nine different discussion-based programs to the development of critical thinking and reading comprehension. Wegerif, Mercer, and Dawes (1999) developed a program they call Talk Lessons, which has been shown to improve quality of discussion as well as problem-solving skills. Reasoning, writing quality, and narrative production have been shown to improve under the influence of collaborative reasoning approaches, which feature student-managed discussion of controversial issues (Reznitskaya et al., 2009), with particularly

strong effects for second language learners (Lin et al., 2012). Other comprehension-focused interventions, such as reciprocal teaching (Palincsar & Brown, 1984), also incorporate discussion, though not always highlighting it as a key component of the program. Interventions to ensure that discussions in math classrooms are issue focused and "accountable" have been shown to improve both math and English language arts (ELA) outcomes (Chapin, O'Connor, & Anderson, 2003). Similar findings have been reported for the value of discussion focused on science in high school (Lemke, 1990), philosophy in middle school (Lipman, 1976), and high school history (Reisman, 2012). Evidence directly relating quantity of discussion in ELA and social studies classrooms to student academic outcomes is strong (Applebee, Langer, Nystrand, & Gamoran, 2003; Gamoran & Nystrand, 1991; Nystrand & Gamoran, 1991; Nystrand, Wu, Gamoran, Zeiser, & Long, 2003).

At the same time, observational studies suggest that rather little discussion goes on in the typical U.S. classroom—less than two minutes per hour on average (Applebee et al., 2003; Gamoran & Nystrand, 1991). The low incidence of discussion is alarming in light of its strong relation to desirable academic outcomes, the importance of authentic language experience for English language learners (e.g., Echevarria, Short, & Powers, 2006; Lesaux, Kieffer, Faller, & Kelley, 2010), and the focus in the Common Core State Standards (Common Core State Standards Initiative, 2010) on oral language and discussion skills as domains in which there are high expectations for student performance.

A formidable challenge, then, is to develop procedures for expanding the amount and improving the quality of classroom discussion. In the study reported here, we evaluated the capacity of a dilemma-focused curricular intervention, Word Generation, to improve the quality of classroom discussion as a mechanism for promoting students' development of general academic vocabulary. Vocabulary instruction has not, to our knowledge, previously been incorporated as a primary goal of discussion-based teaching methods, yet discussion is one approach to meeting conditions known to support vocabulary learning: introduction of target words in semantically rich contexts, opportunities for recurrent exposure to target words, and opportunities for students to use the words in authentic communicative contexts.

Features of High-Quality Classroom Discussion

Classroom discussion is often defined by contrast with teacher monologues or Initiation-Response-Evaluation sequences, rather than by specifying its own defining features. It is variously referred to as exploratory talk (Mercer, Wegerif, & Dawes, 1999), accountable talk (Michaels, O'Connor, & Resnick, 2008), or instructional dialogue (Soter et al., 2008). One study defined good classroom talk as including "peer perspective taking, strong reasoning skills, an ability to connect factual knowledge to the topic, and

an embracing attitude towards newly introduced ideas" (Elizabeth, Ross, Snow, & Selman, 2012, p. 6). A feature of good discussions is high student engagement, as indicated by attentive listening and eagerness to contribute. Good classroom discussions differ from traditional classroom interactions in the participation structure: The ratio of student to teacher talk is high, and students have rights to respond directly to one another. High-quality discussions are also differentiated, though, from many engaged and participatory interactions by the degree of focus: In good discussions, claims, warrants, and conclusions are related to a topic or question (Elizabeth et al., 2012).

Promoting Classroom Discussion

Studies of classroom discussion indicate that certain conversational moves are associated with overall high-quality discussions and in turn, improved student learning. For example, discussions have been found to be more productive when teachers pose authentic, open questions that require reasoning because there is no prespecified answer (Duke, Pearson, Strachan, & Billman, 2011; Reznitskaya et al., 2001). Reznitskaya and colleagues (2001) describe the positioning of students and teachers in such discussions as "co-inquirers" promoting student engagement in "exploring complex concepts, improving their judgments, and responding to each other's reasoning" (p. 33). As students grapple with open-ended questions, teacher demands that they explain their thinking and provide evidence for their claims are further associated with academically productive discussions (Hiebert & Wearne, 1993; Michaels et al., 2008; Wolf, Crosson, & Resnick, 2005). Furthermore, peer-to-peer exchanges in which students listen and respond to one another's ideas and make connections between their own reasoning and that of their peers are particularly important aspects of academically productive classroom discussions (Michaels et al., 2008; Reznitskaya et al., 2001; Wolf et al., 2005). Finally, productive classroom discussions are associated with respectful, collaborative environments that enable engagement with rigorous content and active student participation (Matsumura, Slater, & Crosson, 2008).

At the same time, studies of classroom discussion suggest that many factors constrain quality discourse in school settings. Teachers report anxiety about losing control of the classroom, students who talk too much or too little, and the possibility that discussion will take too much time. In an era of extensively specified content standards, content-focused accountability assessments, and pacing guides, discussion may be seen as a luxury. Lack of teacher knowledge about discussion processes can be another obstacle. The practices teachers identify as discussion often turn out instead to be recitation (Alvermann & Hayes, 1989; Larson, 2000). Using discussion productively requires careful attention to establishing norms and teaching students how to participate productively (Larson, 2000). Thus, it is not

surprising that teachers need considerable support to launch and manage classroom discussions (Chapin, O'Connor, & Anderson, 2003).

Whether quality of classroom discussion relates specifically to students' learning of new word meanings is unclear from the extant research linking discussion to student learning outcomes. Yet, there are reasons to believe that high-quality discussions might play a role in promoting the growth of vocabulary, especially of the general academic vocabulary items needed to make arguments, establish claims, and acknowledge others' contributions. Most saliently, discussions of topics like "Should undocumented immigrants qualify for amnesty?" or "Should marijuana use be legalized" offer multiple opportunities for students to hear and use academic words like *undocumented*, *qualify*, *amnesty*, and *legalize*, as well as raising their interest in reading texts about these topics that also use the target vocabulary items. Thus, discussion is seen as a context for supplementing explicit instruction in word meaning with rich opportunities for incidental learning and for practice.

Vocabulary Interventions

It is widely documented that vocabulary is a domain of deficit for many struggling students and that efforts to close the vocabulary gap are of particular importance because of the relation of vocabulary to reading comprehension (Freebody & Anderson, 1983; Mancilla-Martinez & Lesaux, 2010; Snow, Porche, Tabors, & Harris, 2007). There is no dearth of programs focused on vocabulary improvement that have been demonstrated to have measurable, though typically somewhat limited, impacts. In 1986, Stahl and Fairbanks reviewed the effectiveness of a wide array of vocabulary interventions and concluded that the most effective produced learning of about 300 words per academic year. The National Reading Panel (National Institute of Child Health and Human Development [NICHD], 2000) located 47 vocabulary studies that reported reliable results from experimental or quasi-experimental studies. Only a few vocabulary interventions have been designed for use in the post-primary grades of urban schools serving many language minority students (e.g., Carlo et al., 2004; Graves, 2000; Lesaux, Kieffer, Faller, & Kelley, 2010). These effective programs share some key design features. The words to be taught are limited in number and carefully selected to be of high utility; the words are first presented in meaningful texts, not in isolated lists; student-friendly definitions are provided; students are explicitly taught some word analysis tools (e.g., etymology, morphology, cognates).

Experimental research defining the conditions for incremental word learning provides a strong a priori basis for expecting academic discussion to support word learning. Multiple exposures to words used in different contexts help students learn them better than repeated exposures that are not varied (Bolger, Balass, Landen, & Perfetti, 2008), and understanding improves with each novel exposure (McKeown, 1985). Although illustrations

can support students' learning of concrete academic words (Yanguas, 2009) and simple definitions help students consolidate a core meaning or a word (Bolger, Balass, Landen, & Perfetti, 2008), these two strategies are less well suited for inherently abstract and polysemous terms. Scaffolded discussion with opportunities to use these words in building arguments about interesting topics creates conditions known to support word learning.

The Current Study

The vocabulary intervention tested here, Word Generation, was originally developed in the context of a partnership between the Strategic Education Research Partnership (SERP) and the Boston Public Schools (BPS; see www.serpinstitute.org). The partnership goal, formulated in response to the district's identification of middle grades literacy as a persistent problem of practice, was improving reading comprehension by supporting students' academic vocabulary skills. The approach was developed originally in collaboration with BPS practitioners and consultation from the SERP-convened Design Team. Word Generation was first implemented in BPS, where a quasi-experimental study showed small but significant effects on target word learning (Snow, Lawrence, & White, 2009). Analyses showed that word-learning gains were stronger for language minority than English-only students and that despite summer loss in both the treatment and control schools, program effects were still evident even a year after the completion of the program (Lawrence, Capotosto, Branum-Martin, White, & Snow, 2012).

The Word Generation Program

Word Generation is a cross-content academic language program for middle school students. Each week, the program explicitly teaches five new vocabulary words—such as *relevant*, *presume*, and *indicate*—identified as high leverage academic words, meaning that they are likely to be encountered in academic texts in multiple domains and thus be especially important to developing readers (Coxhead, 2000). These words are also inherently useful in dilemma-focused discussions for constructing positions, weighing ideas, and evaluating arguments.

Throughout the week, students read, talk, and write about the week's topic using the vocabulary words in ELA, math, science, and social studies classrooms. Thus, the program is implemented by a grade-level teaching team rather than by a single teacher—a feature that complicates analyses relating classroom-level predictors to student outcomes. The program includes 24 week-long units. Topics range from those of immediate interest to middle school students, for example, "Should you be able to rent a pet?," to ones of greater general civic interest, for example, "Should there be federal funding for stem cell research?" The different activities in each unit are devised for use in each of the content area classrooms and provide

Table 1
Example Sequence of Word Generation Activities During One Week

Day	Class	Task
Monday	English language arts	Establish word meanings Comprehend gist of the passage
Tuesday	Math	Review words and topic of the week Establish math version of word meaning if applicable Discuss multiple choice math problem Discuss open-response math problem if there is time Discuss relation of math problem(s) to the week's topic
Wednesday	Science	Discuss open-response question about thought-experiment Establish science version of word meaning if applicable
Thursday	Social studies	Identify and argue for or against issue positions Establish social studies version of word meaning if applicable
Friday	English language arts	Writing activity

opportunities for academic discussion about the topic from a disciplinary perspective.

Grounded in vocabulary acquisition research, Word Generation also instantiates principles of learning that link discussion to student growth. The content area activities supported students to engage in whole-class and/or small group discussions that required academic language. Each week's sequence of lessons culminated in a writing activity that tasked students to produce a position essay arguing with support for their point of view. In the following, examples of daily activities are presented along with details of the types of discussion or perspective taking they were designed to support, all with the overarching goal of increasing students' capacity to use and understand academic language. (See Table 1 for an outline of the weekly sequence, and see http://wg.serpmedia.org/ for more information about the program and to download materials, all of which are freely available).

In a typical implementation model, an ELA teacher introduces the topic of the week on Monday. Students or the teacher read an orienting passage aloud while all follow the text. Each passage is written at a sixth-grade level and employs the target words and other features of academic language to introduce multiple points of view about a controversial topic. The teacher also reviews the target words and their meanings and can choose from a list of questions provided in the teacher guide to initiate classroom discussion.

The math activity, often carried out on a Tuesday, provides two options for a problem of the week, one requiring more advanced math skills. The calculation required is embedded in a word problem that is relevant to the

week's topic and, when possible, is based on real-life situations. The teacher guide includes example discussion questions. The set-up for both the math problems and the discussion questions uses some or all of the week's target words. The discussion question introduces a mathematically focused subissue within the weekly topic and challenges students to take a position on it and argue for their position using mathematical concepts.

Typically carried out on Wednesday, the science activity is intended to help students practice thinking like a scientist while using the week's focus words. Real or fictional scenarios are described in brief texts that employ the target words. The scenarios lend themselves to being resolved through data collection. A hypothetical experimental protocol is described, along with a fabricated sample data. Students are given space in their notebooks to write responses to queries about whether the hypothesis is adequately supported by the data, describe the evidence that supports their conclusion, and describe how they might design a better experiment. Teachers are advised to lead a discussion, if time permits, about students' ideas for improving the experiment and to emphasize the use of target words in the discussion.

The social studies debate activity is often done on Thursdays. Students are given four possible positions about the week's topic and told to argue their position by providing reasons and evidence to back up their opinion. If students have a different position from one of the four provided, they are encouraged to argue from that point of view instead. On Friday, students are given a brief writing prompt, usually in the ELA classroom. Students are asked to write a response in which they argue a position on the weekly topic, using focus words from the current or past weeks.

Given the design of the program, we expected schools that participated in the Word Generation program would have better classroom discussions. We also expected that students implementing this program would improve more in their knowledge of taught words than students in control schools. Lastly, we expected that students would learn the meaning of target words better in classrooms that had higher quality discussion, namely, that treatment impacts would be mediated by improved classroom discussion. Thus, our research questions are:

Research Question 1: What was the quality of discussion in treatment and control schools? Were there differences in quality across content areas? Did schools that participated in the Word Generation program demonstrate improved classroom discussion?

Research Question 2: Did participating in the Word Generation program impact students' knowledge of the academic words taught, controlling for individual and teaching team level pretest scores? Did participation affect students' general vocabulary knowledge, controlling for individual and teaching team level pretest scores?

Research Question 3: Did improved classroom discussion mediate the impact of the Word Generation program on students' academic vocabulary knowledge?

Procedures

Methods

District Settings

Twenty-eight schools in two districts participated in this randomized trial. Both are large urban districts in different states in the Northeast with high percentages of students who are eligible for free and reduced lunch. Across the sample, the largest percentage of students were African American (85%, 55%), with fewer White (8.0%, 33%) or Latino/a (5.4%, 11%) students. Since we did not get individualized demographic data from one district, we could not use demographic data as an individual-level variable in our analysis. District recruitment started with an invitation to the leadership team to join the study; central office administrators then asked school leaders if they were interested in participating in this research. Schools that agreed to participate recognized that they might be randomized to either a "Phase 1" or "Phase 2" implementation. Phase 1 schools would receive the treatment the following fall and also in the subsequent school year. Phase 2 schools would serve as controls for two years and would be offered the program in the years subsequent to the completion of the randomized trial (there was no charge to schools for using the program in either condition). Once the schools had agreed to participate, we randomized schools to treatment conditions. Thus, Phase 1 schools were treatment schools during the first year of the trial reported here, and Phase 2 schools are control schools; they will be referred to as such in the rest of the article.

Supporting Teachers to Deliver Word Generation

The Word Generation program introduces some challenging instructional routines and ideas that were novel to many of the implementing teachers. In order to support program implementation, we worked with district personnel to select a Word Generation Lead at each implementing school, to be responsible for coordinating the day-to-day aspects of implementation at that site. Leads were either literacy coaches with joint administrative and teaching duties or experienced full-time teachers.

We provided three levels of professional development support to treatment schools. First, leads in treatment schools from each district were invited to participate in a three-day summer institute in Cambridge, Massachusetts. During the institute, participants took part in sessions addressing the rationale and research base of the Word Generation program. They attended sessions by researchers and practitioners about three topics: a discussion framework based on Accountable Talk (Michaels et al., 2008), the importance of cross-disciplinary collaboration as a means of supporting school internal coherence (Elmore, Forman, Stosich, & Bocala, 2013), and how to support

academic language across the content areas. They worked together to preview the Word Generation program and analyze examples of lessons. Leads from 11 of 15 treatment schools attended the summer institute.

Second, leads and all teacher participants in treatment schools were invited to attend half-day, comprehensive sessions on the Word Generation program in their respective districts in late summer. These sessions were co-led by Word Generation's primary content developer together with an expert on the discussion component of the program. During these sessions, participants learned about the rationale for Word Generation, its goals and structure, and key components such as teacher talk moves effective in promoting discussions across the content areas. Participants viewed video examples of Word Generation lessons in each of the content areas. Although all teacher participants in each district were invited to attend these sessions, other commitments scheduled by the districts made it difficult for teachers to attend. The majority of schools sent two to three teachers to these sessions. One school did not send any representatives.

Third, all treatment schools were offered professional development on an individualized basis in the form of on-site meetings with study staff during the course of the school year. The foci of these meetings ranged from introducing participating teachers to Word Generation to resolving specific questions about implementation. Study staff corresponded regularly (no less than monthly) with the leads from each school to check in and see if any additional supports were desired. Upon request, meetings with grade-level teams were held to answer questions, provide additional information about Word Generation, or model lessons.

The professional development supports described previously and Word Generation curricular materials were offered to control schools after the trial phase of the study was concluded.

School Settings

Table 2 presents some basic information about each of the 28 participating schools. The first and second data columns provide the school and district code, respectively. The third data column provides the total enrollment for each school across all grades. On average, enrollment was higher in the treatment (M = 486) than it was in the control schools (M = 396). Unfortunately, the districts were located in different states, each with its own achievement test. The proficiency level for the Maryland Grade 8 state test was equivalent to a scaled score of 239 on the 2009 NAEP, while the proficiency threshold for the Pennsylvania Grade 8 state test was equivalent to 245 on NAEP. Thus, it is clear that the proficiency cut-off was somewhat more stringent in Pennsylvania, a situation that probably holds for the sixth- and seventh-grade tests as well (though NAEP scaled scores are unavailable for these grades). Fortunately, though, both state tests provided proficiency

benchmarks that we could use to establish the average proficiency in sixth, seventh, and eighth grade for each school. On average, control schools had higher baseline proficiency than treatment schools within both states. All students in participating grades/schools were invited to participate in the study, and only those whose parents granted informed consent were included in the analysis.

Unfortunately, not all these students contributed to the analysis. Although we collected pretest data from 3,754 students, many schools received their posttests later than scheduled. Since eighth-grade graduation ceremonies were underway in many schools when the posttests were delivered, only 28% of eighth-grade students who completed the pretest completed a posttest. Although sixth-grade and seventh-grade completion rates were higher (45% and 54%, respectively), they were still disappointingly low. Completion rates varied somewhat across treatment (49%) and control conditions (35%) and were higher in the smaller of the two districts (77%) and lower (38%) in our larger partner district, which had more schools to distribute the tests to. We have no reason to believe that differences in testing completion rates were nonrandom; they reflected logistical difficulties with our vendor. We tested to see if there were pretest differences between students who did (M = 18.59) and did not (M = 18.93) complete both waves of academic vocabulary testing; there were no differences, $t(3,752) = 1.595$, $p = ns$.[1] Table 2 indicates the number of students at each grade level who contributed both pre- and posttest data to this study. In total, 622 students from 13 control schools contributed pre- and posttest data, and 932 students from 15 treatment schools contributed pre- and posttest data. Across all schools, 8 contributed data from all grade levels, 17 schools contributed data from two grade levels, and 3 schools contributed data from only one grade level.

As the description of the program makes clear, full implementation requires participation by a team of math, science, social studies, and English teachers. We believe that implementation across cross-content areas is one of the strengths of the program; however, it does present some challenges for program evaluation. The first research question asks about the impact of the program on classroom discussion. We were able to conduct 168 observations across randomly sampled content-area classes. This meant that we were able to calculate effect sizes with observation-level data. Unfortunately, observations were too sparse to calculate reliable team-level estimates of classroom discussion quality, since we know that there are large average differences in average baseline discussion levels in each content area. Thus, if in one teaching team only the math or science class was observed, we could not generalize from that observation to the rest of the team. Therefore, while we calculate discussion treatment effects with observation-level data, we use school-level aggregated scores in the mediation models used to answer Research Question 3 (these are explained in more detail in the

Table 2

School Demographics and Grade-Level Contributions for Each School by Treatment Conditions

Treatment Condition	School Code	District	Total Enrollment	Percent Proficiency (Sixth Grade)	Percent Proficiency (Seventh Grade)	Percent Proficiency (Eighth Grade)	Percent Free and Reduced Lunch	Percent English Language Learner	Grade 6 Contributions	Grade 7 Contributions	Grade 8 Contributions	Total Contributions
Control	1	1	286	64	76	56	94	0	11	17	0	28
schools	3	1	442	80	73	73	84	0	24	15	0	39
	5	1	515	62	69	NA	81	0	24	19	0	43
	6	1	237	90	86	86	82	0	38	17	0	55
	11	1	273	76	82	70	96	75	17	25	15	57
	13	1	966	71	60	54	86	18	27	43	59	129
	20	1	387	87	93	74	88	0	11	12	0	23
	35	1	269	57	47	44	86	0	21	14	17	52
	36	1	647	98	82	85	74	0	69	74	0	143
	30	2	297	31	45	75	95	0	7	0	0	7
	31	2	329	35	19	39	92	0	6	4	0	10
	32	2	286	46	67	81	55	0	12	0	0	12
	33	2	216	54	65	72	88	0	14	10	0	24
Average			396.2	65	67	67	85	7	21.6	19.2	7.0	47.8
Sum			5,150						281	250	91	622
Word	8	1	405	90	96	81	96	0	0	29	0	29
Generation	9	1	442	85	68	71	82	0	39	28	11	78
schools	10	1	452	73	32	58	79	0	41	41	26	108
	12	1	505	79	82	56	77	14	32	33	23	88
	15	1	817	80	79	78	81	0	0	60	0	60
	16	1	501	64	69	63	89	0	54	87	115	256
	17	1	328	53	32	41	96	0	5	0	0	5
	18	1	494	80	77	64	81	0	43	29	0	72
	21	2	680	24	24	43	94	0	0	5	0	5
	22	2	680	73	31	69	93	0	0	0	12	12
	23	2	368	32	52	58	91	0	18	13	17	48
	24	2	529	28	37	61	92	0	0	0	15	15
	25	2	317	26	36	55	91	10	0	45	0	45
	26	2	456	51	59	77	77	54	0	60	0	60
	27	2	321	60	73	82	74	0	0	0	51	51
Average			486.3	60	56	64	86	5	15.5	28.7	18.0	62.1
Sum			7,295						232	430	270	932

following). The second research question examines the impact of treatment on individual scores in a multilevel context that accounts for student- and team-level pretest scores.[2]

Measures

Treatment. TREAT is a bivariate variable that specifies if a teaching team participated in the Word Generation program (TREAT = 1) or not (TREAT = 0). Treat is a Level 2 predictor in some of our analyses.

Discussion Quality Measures

A classroom observation rubric was developed to measure several dimensions of discussion quality (each dimension is described in the following). Rater training was led by site directors (who were members of study staff) in their respective school districts. Site directors first achieved interrater reliability by rating videos of Word Generation lessons and debriefing scores for two lessons per content area. Site directors then conducted live observations of both Word Generation lessons and "business as usual" instruction in six classrooms in an urban school district that was not participating in the study and developed an extensive coding manual to support the rubric descriptions. Interrater reliability was considered acceptable when the two raters had achieved at least 90% within one-point agreement for three consecutive observations.

Each site director then assembled a team of graduate students in education to serve as raters in their respective districts. Three raters in each district participated in a full-day training session in which they were introduced to the study and the observation instrument. For each rubric, raters were taught to identify observable indicators for making scoring decisions. Indicators were selected to be as low inference as possible. Photos, video clips, and transcripts were analyzed to illustrate each score point for each rubric. Raters were instructed to sketch a classroom map to track participation rates and to transcribe the lesson talk as thoroughly as possible; raters were instructed to use standardized codes for teacher and student conversational moves. The training session concluded with mock observations of three video clips from math and English language arts lessons. For each mock observation, raters were required to record evidence of the observable indicators, score all rubrics, and justify scores with evidence. Though the presence of Word Generation curricular materials and topics in the treatment classrooms precluded blinding observers to condition, hypotheses about the role of discussion quality in Word Generation were not shared with them. Furthermore, as noted previously, the indicators used in applying the rubric were observable and low inference.

Interrater reliability training for the teams of raters in each district was carried out via live observations in schools over the following five school days.

Training observations were conducted only in classrooms that had not been sampled for actual observation in the data collection period. All three raters-in-training and the site directors attended six training observations. Training was concluded when all raters-in-training had achieved 90% within one-point agreement with the site director for three consecutive observations.

Classroom observations were scheduled through a three-step process. First, within each content area and school, participating teachers were randomized, and a schedule was created identifying specific teachers and classrooms to be visited on given day. Second, schools were notified of the first choice teachers and alternates and asked to confirm whether, in treatment schools, the identified teacher would be teaching Word Generation that day and in control schools whether regular instruction (e.g., as opposed to testing or independent student work) was planned. Third, adjustments were made to the schedule based on school-provided information, and schools were given the final schedule. We ended up visiting most schools in February or March and then again in May. In total, we conducted 168 observations of content-area classrooms across the two districts. Of these, 44 observations were conducted with two raters, and one was conducted with three raters to ensure that the high interrater reliability we established in training was maintained in the field (information on interrater reliability is provided for each rubric category in the following). Extensive notes and transcriptions were taken at all observations (examples are provided in the following), but the ratings themselves were completed immediately after the completion of each observation. We found that the four rubric categories related to quality of classroom discussion were moderately correlated (r ranged from .56 to .77; Table 3). The four dimensions are as follows.

Support for participation. Each observer rated to what extent students were engaged in classroom discussion and if the teacher created a well-ordered and respectful environment that enabled engagement with lesson content and participation in the discussion, as there is evidence that respectful, prosocial environments are essential prerequisites to rich discussions (Matsumura et al., 2008); these ratings were used to create the PARTICIPATION variable. Classrooms were rated on a 3-point scale where a rating of 1 was reserved for classrooms characterized by student hostility, widespread lack of participation, or chaos, and a rating of 3 indicated that nearly all students appeared consistently engaged with minimal side talk or distractions. Interrater reliability was high (r = .93, with 41 of the 45 observations getting exact matches in this category).

Student engagement. ENGAGEMENT specifies the percentage of students participating in or attending to the classroom discussion. Rate of student participation is fundamentally important to enabling rich discussions, and rates of student talk have long been of interest in studies examining

Table 3
Summary of Intercorrelations of the Four Rubric Categories Related to Quality of Classroom Discussion

		1	2	3	4
1.	Support for participation	—			
2.	Student engagement	.77***	—		
3.	Teacher talk moves	.56***	.63***	—	
4.	Substantive contributions	.60***	.61***	.66***	—

***$p < .001$.

the relationship between student talk and learning outcomes (Murphy et al., 2009). A rating of 1 indicated that a quarter or fewer of the students participated in discussion during the observation period, whereas the maximum score (ENGAGEMENT = 3) was awarded when 50% to 100% of students participated in discussion during the observation period. Interrater reliability was high (r = .90, with 39 of the 45 observations getting exact matches in this category).

Teacher talk moves. The teachers' ability to facilitate high-quality discussion was marked by the occurrence of open-ended questions, follow-up questions requiring students to explain their thinking, and other teacher talk moves that extended participation and perspective taking in whole-class discussion. There is evidence that authentic, open-ended questions that require reasoning and evidence are central to academically productive discussions (Michaels et al., 2008; Reznitskaya et al., 2009; Wolf et al., 2005). The lowest score (TALK_MOVE = 1) was given to classroom discussions in which all teacher questions had single, known answers (closed questions). The highest ratings (TALK_MOVE = 5) were reserved for classrooms where the teacher initiated a range of question types including open-ended questions and also asked students to provide evidence or explain their ideas more clearly. Interrater reliability was high (r = .96, with 40 of the 45 observations getting exact matches in this category).

Substantive contributions. SUBSTANTIVE rates the intellectual contributions of students during classroom discussions. The lowest ratings (SUBSTANTIVE = 1) indicates classrooms in which students provided only perfunctory answers. The highest ratings (SUBSTANTIVE = 5) were reserved for discussions in which multiple students elaborated ideas and explained their thinking while providing evidence. In these classrooms, students also asked each other to explain their thinking or explicitly linked their own to others' contributions. Such peer-to-peer exchanges in which students respond to one another's ideas are an important element of rich discussion (Michaels

et al., 2008; Reznitskaya et al., 2001). Interrater reliability was high (r = .93, with 39 of the 45 observations getting exact matches in this category).

Composite discussion quality rating. To establish overall discussion quality ratings for each observed classroom, we first divided each observation rating by the number of possible points available in that category (to put the 3-point and 5-point ratings on the same scale) and tallied the sum of those scores for a maximum of 4 points (COMPOSITE = PARTICIPATION/3 + ENGAGEMENT/3 + TALK_MOVE/5 + SUBSTANTIVE/5). We also created scores based on factor weightings for the four variables but found that scores derived by the two methods were so highly correlated (r = .97) that there was no reason to use the less transparent factor-weighted scores. Composite discussion quality rating ranged from 1.06 to 3.60 out of a possible total of 4.00.

Standardized discussion quality rating. Because average COMPOSITE discussion scores varied systematically across content area, we needed a standardized rating for each content area for use in a school-level quality rating. We transformed the COMPOSITE scores of each content area to z-scores. So, for instance, the average zCOMPOSITE score of all 46 observations of ELA classes is 0, with a standard deviation of 1 (and the same holds true for the other content areas). Although we did not use these scores directly in our analysis, we used the average zCOMPOSITE score for the observations conducted at each school to create the weighted school-level discussion quality rating (QUALITY), which we describe next.

Weighted school-level discussion quality ratings. We created the school-level discussion quality ratings for each school by averaging the zCOMPOSITE scores for the observations conducted at the school. In this way, schools that happen to have had more math observations then ELA observations would not be penalized due to the fact that average discussion quality was lower in math classes.[3] The average of QUALITY scores across all the schools is 0, and the standard deviation = 1.

Assessments

Academic Vocabulary

Academic Vocabulary Knowledge (individual score). We measured student knowledge of taught academic vocabulary with a 36-item multiple-choice curriculum-based test to create the outcome variable ACA_VOC_W2. Each assessment item presented a target word, underlined, in a neutral sentence context; the students chose from among four options the one that was the closest synonym for the target word. Although synonym items do not provide much scope for students to demonstrate partial knowledge of target words, it is probably the most common kind of vocabulary

assessment item; we used them to facilitate comparison with our previous studies and other vocabulary intervention research. Target words were taken from the Word Generation curriculum; for the most part, they are found on the Academic Word List (Coxhead, 2000). We used the student total score on this assessment as an outcome in our analysis; scores ranged from 1 to 36 at pretest. The scale reliability coefficient was acceptably high for these 36 items (Cronbach's $\alpha = .79$).

Academic vocabulary knowledge (teaching team mean). Mean vocabulary scores for each grade level teaching team in each school were used to create a Level 2 pretest covariate ACA_VOC_TTM_W1. This variable was created using only the scores of students who completed both pre- and posttests. ACA_VOC_TTM_W1 scores ranged from 12.15 to 24.54.

Academic vocabulary knowledge (teaching team mean centered). Teaching team mean centered scores (ACA_VOC_TTMC_W1) were created by finding the difference between the individual score of each student and the mean score of the teaching team in which he or she was placed (ACA_VOC_TTMC_W1 = ACA_VOC_W2 – ACA_VOC_TTM_W1). Scores were created using data from students who completed both waves of data collection and ranged from –20.75 to 15.08. ACA_VOC_TTMC_W1 was used as a Level 1 pretest covariate in our analysis.

General Vocabulary

General vocabulary knowledge (individual score). Depending on the grade level of the student, the extended scale scores from Level 6 or Level 7/9 of the Gates-MacGinitie vocabulary assessment at posttest were used to create VOCAB_W2. The extended scaled scores were developed according to item response theory using the Rasch model (MacGinitie, MacGinitie, Maria, & Dreyer, 2000). This assessment presented students with a sentence or clause with an underlined target word. Students were required to select a synonym for the underlined word from five options. The words assessed in this 45-item battery include very frequent words, general academic words, and also very rarely used words. Test reliability was high in our sample (.87), with extended scale scores ranging from 367 to 661 at pretest. VOCAB_W2 was analyzed as an outcome in our analysis.

General vocabulary knowledge (teaching team mean). We used mean Gates-MacGinitie extended scale scores for each grade-level teaching team in each school at pretest to create the variable VOCAB_TTM_W1. This average only included students who completed both pre- and posttest, with teaching team average scores ranging from 443.66 to 552.5. VOCAB_TTM_W1 is used as a Level 2 pretest covariate in our analysis.

General vocabulary knowledge (teaching team mean centered). Teaching team centered individual Gates-MacGinitie extended scale scores were created for each student by finding the difference between each student's score and the average score of the teaching team that he or she was assigned to (VOCAB_TTMC_W1 = VOCAB_W1 – VOCAB_TTM_W1). Scores ranged from –19.39 to 23.21. VOCAB_TTMC_W1 is used as a Level 1 pretest covariate in our analysis.

Covariates

Grade level. The district provided us with information about students' grade level. We used these data to create two variables. GRADE6 indicates if a student is in sixth grade (GRADE6 = 1) or not (GRADE6 = 0). GRADE7 indicates if a student is in seventh grade (GRADE7 = 1) or not (GRADE7 = 0). These variables were used as Level 1 covariates.

Grade-level proficiency scores. The two districts that participated in this study were in different states, and each state uses its own assessment to determine student reading proficiency. It was impossible for us to scale across the two state achievement measures. Instead, we used the percentage of students who reached proficiency by local state standards in this analysis at each grade level in each school as a Level 2 covariate. Percentage of students who scored proficient ranged from quite low to quite high at sixth (6G_PPROF = .11 to 6G_PPROF = .98), seventh (7G_PPROF = .13 to 7G_PPROF = .96), and eighth grade (8G_PPROF = .19 to 8G_PPROF = .89). We used these scores as Level 2 covariates.

School percentage free and reduced lunch scores. We used publicly available data from online sources to create PERCENT_FARM, a variable that records the percentage of students eligible to participate in the federal free and reduced lunch program at each school and was used as a Level 2 covariate. PERCENT_FARM values ranged from moderate (PERCENT_FARM = 54.9) to quite high (PERCENT_FARM = 95.6) in our sample of urban schools.

School percentage special education. We established the percentage of students who are on individual educational plans from publicly available sources and used these data to create PERCENT_SPED, which we used as a Level 2 covariate in our analysis. School-level values on this variable ranged widely across our sample from PERCENT_SPED = 7.1 to PERCENT_SPED = 36.2.

School district. DISTRICT is a bivariate variable used to specify if a student was in District 1 (DISTRICT = 0) or District 2 (DISTRICT = 1).

Analysis

Within each of the participating districts, schools were randomly assigned to implement the Word Generation program or to a "treatment as usual" control condition. Schools were the unit of random assignment, rather than classrooms within schools, because Word Generation is implemented by cross-disciplinary teaching teams; in many schools, there is only one teaching team per grade, precluding within-school comparisons. Furthermore, within-school comparisons are more likely to produce contamination of the control condition. Nonetheless, the choice of school as the unit of randomization had some drawbacks: School leaders agreed to participate, but the teachers within schools varied in their level of commitment, and schools were differentially subject to disruptions (e.g., influx of new student groups, transitions in leadership) and factors (e.g., negative school climate, poor physical conditions) that would have been held constant in comparing classrooms within schools.

Schools were first ranked within district on a school-level covariate composite (percentage minority, percentage free and reduced lunch, percentage English language learners, and prior mean achievement using the state accountability assessment). Then schools were grouped into blocks of two based on their composite score and randomly assigned to treatment or control within each block to maximize comparability of treatment and control schools. This strategy has been recommended to minimize group differences and reduce the potential for unhappy randomization when the number of units to be randomized into groups is small. Some of these school-level variables were also used as covariates in the analysis, which serves to reduce the intraclass correlation (ICC) and thereby increase power for detecting group differences (Bloom, Richburg-Hayes, & Black, 2007).

Testing Effects on Discussion

We examine differences between the quality of discussion in treatment and control schools by comparing composite discussion quality rating scores in the two sets of schools and test differences with two-sample mean comparison *t* tests. We determine overall effect sizes as well as treatment effects in each content area. As a step in our mediation analysis we also fit a random effects model on weighted school-level discussion quality ratings (described in more detail in the following).

Testing Effects on Vocabulary

Grade-level teaching teams were the unit of implementation in each school,[4] and thus the equivalent of "teacher" or "classroom" in studies that do not use cross–content area designs. We explored intention to treat (ITT) effects with the following Level 1 equation:

$$\text{ACA\_VOC\_W2} = \beta_{0j} + \beta_{02}\text{ACA\_VOC\_TTMC\_}W1_{ij} + \beta_{03}\text{GRADE6}_{ij} + \beta_{04}\text{GRADE7}_{ij} + \varepsilon_{ij}, \quad (1)$$

where ACA_VOC_W2_{ij} is the predicted posttest academic vocabulary score of child i in classroom j; $\beta_{02}\text{ACA\_VOC\_TTMC}_{ij}$ represents the difference in the predicted posttest vocabulary score associated with differences in student pretest teaching team centered vocabulary; $\beta_{03}\text{GRADE6}_{ij}$ and $\beta_{04}\text{GRADE7}_{ij}$ represent the differences in posttest scores associated with student grade level controlling for other pretest covariates; ε_{ij} is the residual error term. β_{01} is the adjusted teaching team level average vocabulary score defined by

$$\beta_{0j} = \gamma_{01}\text{ACA\_VOC\_TTM\_W1}_j + \gamma_{02}\text{TREAT}_j + \gamma_{03}6\text{G\_PROF}_j + \gamma_{04}7\text{G\_PROF}_j + \gamma_{05}\text{PERCENT\_FARM}_j + \gamma_{06}\text{PERCENT\_SPED}_j + \gamma_{07}\text{DISTRICT}_j + \upsilon_j, \quad (2)$$

where β_{0j} is the adjusted teaching team average vocabulary score for team j; $\gamma_{01}\text{ACA\_VOC\_TTM\_W1}_j$ is the predicted difference in posttest associated with teaching team average pretest vocabulary; $\gamma_{02}\text{TREAT}_j$ is the estimated difference in posttest vocabulary between teams that did and did not implement Word Generation; $\gamma_{03}6\text{G\_PROF}_j$ and $\gamma_{04}7\text{G\_PROF}_j$ represent differences in predicted vocabulary associated with the a one-point difference in percentage of proficient students at Grades 6 and 7; $\gamma_{05}\text{PERCENT\_FARM}_j$ estimates the difference in vocabulary posttest scores associated with a one-point difference in the percentage of students at the school who qualify for the federal free and reduced lunch program;$\gamma_{06}\text{PERCENT\_SPED}_j$ paramerizes the predicted difference in vocabulary associated with a one-point difference in the percentage of students with individualized education plans at each school; and $\gamma_{07}\text{DISTRICT}_j$ represents the difference between predicted posttest scores by district, controlling for all other predictors; υ_j captures the unexplained variance at the second level of the model.

Overview of Analytic Plan to Explore How Discussion Mediates Program Treatment

We wished to determine if improved discussion mediated the hypothesized relationship between treatment and vocabulary outcomes. This analysis has been described as a 2→2→1 approach, since the treatment and mediator variables are measured at Level 2, but the outcome is a Level 1 variable (Krull & MacKinnon, 2001). Several approaches have been used for modeling mediation analysis in non-nested mediational contexts. One traditional approach estimates treatment effects on the student-level outcome (*c path*) and then replicates that analysis including the mediator (*c prime*

path). The difference between the parameter estimates of treatment in the full model and the direct treatment is the mediated effect (Judd & Kenny, 1981). A second approach analyzes the indirect effect as the product of the treatment coefficient in a model predicting the mediator (*a path*) and the estimated impact of the mediator on the outcome variable (*b path*) controlling for treatment. In single-level contexts, the two approaches have been shown to be algebraically identical (MacKinnon, Warsi, & Dwyer, 1995). Bootstrapping is a particularly flexible and powerful approach to establishing the significance and confidence intervals of indirect effects since they are based on empirical estimation of the sampling distribution of the indirect effect rather than assumptions of normality (Efron & Tibshirani, 1993). "Bootstrapping provides the most powerful and reasonable method of obtaining confidence limits for specific indirect effects under most conditions, so our primary recommendation is to use bootstrapping—in particular, BC bootstrapping— whenever possible" (Preacher & Hayes, 2008, p. 886).

Results

Research Question 1: What was the quality of discussion in treatment and control schools? Were there differences in quality across content areas? Did schools that participated in the Word Generation program demonstrate improved classroom discussion?

Table 4 presents the observation scores in treatment and control schools by content area. In control schools, discussion scores in math and science classes tended to be about half a standard deviation lower than scores in English and social studies classes (Column 1). In the treatment schools (Column 2), on the other hand, math classes had the highest discussion score on average, with science classes still lagging behind social studies and English. Word Generation classes in each content area had higher average ratings for classroom discussion than control classes. Differences between treatment and control classes were particularly strong for math (Cohen's d = 1.13) and much smaller for ELA (Cohen's d = 0.44).

Figure 1 presents a histogram depicting the discussion rating frequencies in the control and the treatment schools. Compared to the control schools, there were far fewer classrooms in treatment schools with very poor ratings, indicating low levels of participation, few open-ended questions, and little student engagement. Classrooms in Word Generation schools were more frequently rated at the medium to high level of discussion quality.

Research Question 2: Did participating in the Word Generation program impact students' knowledge of the academic words taught, controlling for individual and teaching team level pretest scores? Did participation affect students' general vocabulary knowledge, controlling for individual and teaching team level pretest scores?

Table 4

Composite Discussion Quality Rating by Treatment Status and Content Area Observed With Estimated Effect Sizes

Content Area Observed	Control Schools		Word Generation Schools		Total		Pooled Standard Deviation	Difference (Word Generation – Control Schools)	Effect Size (Cohen's *d*)
	Mean	*n*	Mean	*n*	Mean	*n*			
Math (*n* = 42)	1.88	23	2.84	19	2.31	42	0.85	0.96	1.13
	(0.83)		(0.87)		(.97)				
Science (*n* = 36)	1.85	21	2.27	15	2.02	36	0.91	0.42	0.47
	(0.90)		(0.92)		(0.92)				
Social studies (*n* = 44)	2.29	19	2.61	25	2.48	44	0.84	0.32	0.38
	(0.84)		(0.84)		(0.84)				
English language arts (*n* = 46)	2.30	25	2.65	21	2.46	46	0.80	0.35	0.44
	(0.85)		(0.73)		(0.81)				
Total (*n* = 168)	2.08	88	2.61	80	2.33	168	0.86	0.53	0.62
	(0.87)		(0.84)		(0.89)				

Note. Mean with standard deviations in parentheses.

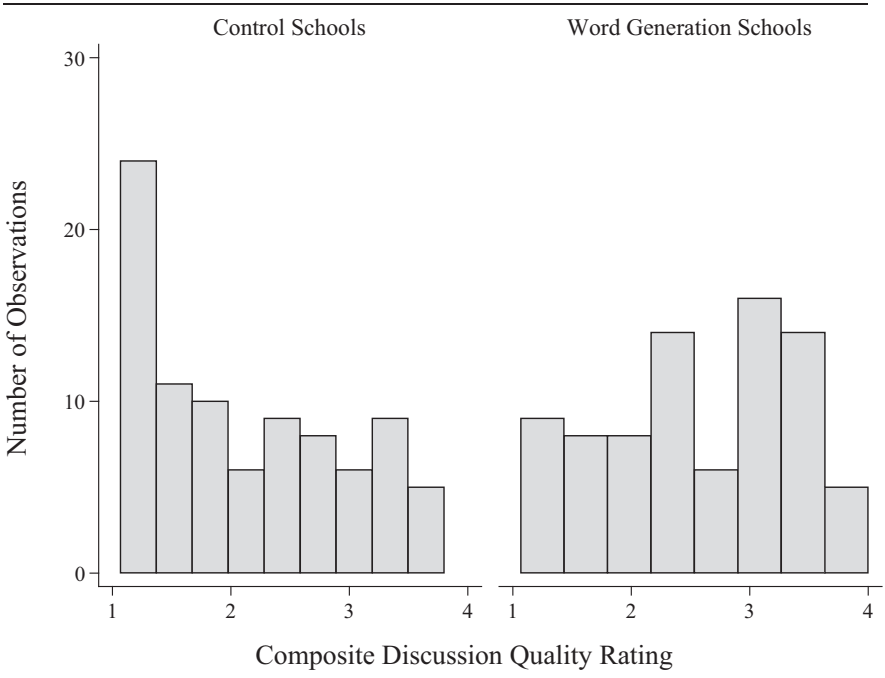

Figure 1. **Histogram of composite discussion quality rating given in classroom observations by treatment status.**

Table 5 presents overall pretest and posttest scores for students in Word Generation and control schools. Despite the careful randomization process, students in treatment schools had higher vocabulary knowledge at pretest. Therefore we used differences between students' posttest and pretest scores in the treatment and control schools to calculate preliminary effect sizes.[5] We found that average improvement in the treatment schools ($\delta T = 1.74$) was 1 point greater than average improvement in the control schools ($\delta C = 0.71$), suggesting an average treatment effect of 0.17 (using the pooled standard deviation; $\sigma = 6.14$).[6] The Gates-MacGinitie Vocabulary pretest also favored the treatment schools; however, we found that improvement in the comparison schools ($\delta C = 6.26$) was actually higher than in the treatment schools ($\delta T = 4.98$), yielding an average treatment effect of –0.04 (using the pooled standard deviation; $\sigma = 32.37$). These estimates are very rough. They ignore known individual- and school-level covariates and are calculated at the student level even though teaching team was the unit of implementation.

These estimates of effect size ignore the fact that the Word Generation program is actually administered by grade-level cohorts of teachers in

Table 5

Academic and General Vocabulary Test Scores by Treatment Status With Estimated Effect Sizes

	Overall Sample Mean			Control Sample Mean				Word Generation Sample Mean				Pooled Standard Deviation	$\delta T - \delta C$	Effect Size Calculated From Raw Scores	Difference Calculated From HLM	Effect Size Calculated From HLM	p Value
	Pretest	Post test	N	Pretest	Post test	n	δC	Pretest	Post test	n	δT						
Academic vocabulary	18.57 (6.15)	19.89 (7.10)	1,554	18.05 (6.01)	18.76 (6.86)	622	0.71	18.91 (6.22)	20.65 (7.15)	932	1.74	6.14	1.03	0.17	1.53	0.25	<0.01
General vocabulary	505.29 (32.37)	510.79 (34.91)	1,416	502.11 (33.41)	508.38 (36.40)	575	6.26	507.45 (31.47)	512.44 (33.78)	841	4.98	32.27	−1.28	−0.04	−0.38	−0.01	*ns*

Note. Academic vocabulary measured by the Word Generation multiple choice test; general vocabulary measured with the extended scale scores from Level 6 or Level 7/9 of the Gates-MacGinitie vocabulary assessment. Standard deviations in parentheses. HLM = hierarchical linear modeling.

schools that vary on many dimensions. Figure 2 presents box plots showing that Word Generation schools started the study with slightly higher vocabulary scores and that District 2 had higher pretest scores than District 1. In order to determine if treatment effects were statistically significant when accounting for the nesting of students within teaching teams, we fit a series of multilevel models presented in Table 6. Model A predicts academic vocabulary as an individual-level outcome from the mean of the teaching team, the academic vocabulary of each individual centered on the mean of their cohort, and treatment status. Model B is similar except that we include school- and cohort-level covariates as well as school district as a fixed effect. We did not find interactions between treatment and covariates. Treatment effects were similar in models with ($TREAT_j = 1.529, p < .01$) or without covariates ($TREAT_j = 1.264, p < .05$). Model C and Model D are similar except that they predict students' vocabulary as measured on the Gates-MacGinitie Vocabulary Assessment. These models show no significant treatment effect when we account for pretest vocabulary scores ($TREAT_j = 0.0151, p = ns$) nor when we also account for other covariates ($TREAT_j = -0.375, p = ns$). In order to calculate effect sizes with these more accurate estimates of the impact of treatment calculated at the teaching team level, we need to use the standard pooled deviation, not the standard deviation calculated at the group level (What Works Clearinghouse, 2008). We present these on the right side of Table 5. Our best estimate of the effect of treatment on academic words learned is Hedge's $g = 0.25$, $p < .01$.

Research Question 3: Did improved classroom discussion mediate the impact of the Word Generation program on students' academic vocabulary knowledge?

Table 7 presents the parameter estimates that we need to conduct mediation analysis. The first column presents a model predicting academic vocabulary from treatment controlling for pretest vocabulary. In traditional mediation analysis, this is known as the *c prime path*. The first column provides an estimate of treatment on composite discussion quality ratings ($TREAT_j = 0.393, p < .05$) controlling for pretest scores. The third column provides an estimate of composite discussion quality on academic vocabulary controlling for pretest scores and treatment ($DISCUSSION_j = 0.516, p = ns$). The *indirect effect* is equal to the product of the estimated treatment impact on classroom discussion and the estimated impact of classroom discussion on student vocabulary outcomes (*indirect effect* = .393 × .516 = .203). Since we know (from the third column in Table 7) what the direct effect of treatment controlling for pretest and discussion is ($TREAT_j = 1.18, p < .05$), we can calculate the total effect (*total effect* = *indirect effect* + *direct effect* = 0.203 + 1.18 = 1.38). This estimate is similar to that which we obtained in the hierarchical linear modeling (HLM) models including covariate controls ($TREAT_j = 1.529, p < .01$). We established confidence intervals for the direct,

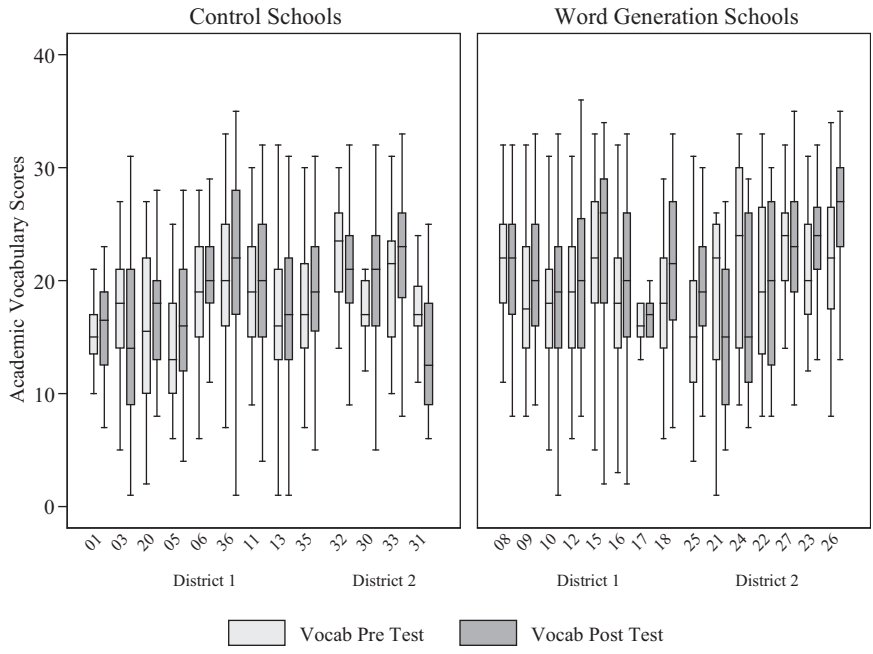

Figure 2. **Boxplot of academic vocabulary scores for each school by treatment status.**

indirect, and total effect from a derived bootstrapped sampling distribution of the indirect path (the product of the *a path* and *b path*; Table 8). These confidence intervals suggest that about 14% of the total effect of Word Generation is mediated through improved discussion ($p < .05$).

The classrooms in which active, engaged discussion took place looked very different from those in which more traditional recitation was the norm. Raters were trained to transcribe as closely as possible all contributions by teachers and students during lesson observations. Raters then used these transcriptions as evidence for rating decisions made immediately following the observation. Transcription notes were not revised in any way. We selected the following examples based on two criteria. First, as it was not possible to transcribe all teacher and student talk, we selected transcriptions that were most complete and legible. Second, since our purpose was to illustrate some possible mechanisms by which high-quality discussions might support academic word learning, we selected observations that had received high ratings. Consider the following two extracts from observation notes.

Table 6

Hierarchical Linear Models Predicting Students' Academic and General Vocabulary Scores From Pretest Scores (Models A and C) and From Pretest Scores Controlling for School-Level Covariates and Student Grade Level (Models B and D)

Outcome	Model A Academic Vocabulary	Model B Academic Vocabulary	Model C General Vocabulary	Model D General Vocabulary
Treatment (TREAT)	1.264*	1.529**	0.015	−0.376
	(0.500)	(0.487)	(2.323)	(2.168)
Academic vocabulary teaching team mean (ACA_VOC_TTM_W1)	0.837***	0.808***		
	(0.097)	(0.142)		
Academic vocabulary team mean centered (ACA_VOC_TTMC_W1)	0.697***	0.697***		
	(0.024)	(0.024)		
General vocabulary teaching team mean (VOCAB_TTM_Wl)			0.691***	0.582***
			(0.070)	(0.090)
General vocabulary team mean centered (VOCAB_TTMC_Wl)			0.798***	0.798***
			(0.022)	(0.022)
Sixth-grade percentage proficient (6G_PPROF)		−2.112		−0.061
		(2.750)		(12.220)
Seventh-grade percentage proficient (7G_PPROF)		3.796*		8.552
		(1.711)		(7.573)
Percentage free and reduced lunch (PERCENT_FARM)		−0.014		−0.205
		(0.036)		(0.161)
Percentage special education (PERCENT_SPED)		−0.004		0.052
		(0.040)		(0.178)
Sixth grade (GRADE6)		0.546		−0.155
		(0.811)		(3.147)
Seventh grade (GRADE7)		0.981		−0.204
		(0.636)		(2.710)
District 1 (DISTRICT)		0.078		−10.24*
		(1.078)		(4.710)
Intercept	3.397	3.387	160.9***	235.0***
	(1.757)	(4.999)	(34.920)	(50.430)
Level 2 variance (teaching team)	28.29***	28.25***	530.7***	530.7***
	(0.517)	(0.516)	(10.180)	(10.180)
Residual	1.775	1.406	40.88***	27.25***
	(0.313)	(0.267)	(6.785)	(5.361)
N	1,554	1,554	1,416	1,416
−2 log likelihood	9,654.227	9,645.341	12,956.545	12,944.509

Note. Standard errors in parentheses. Academic vocabulary measured by the Word Generation multiple choice test; general vocabulary measured with the extended scale scores from Level 6 or Level 7/9 of the Gates-MacGinitie vocabulary assessment.
*$p < .05$. **$p < .01$. ***$p < .001$.

Example 1:
Topic: Single-gender education, Unit 12
Math
Target Words: *paradigm*, *comprise*, *gender*, *conduct*, *adapt*

Table 7
Estimates of A, C, and C Prime Paths Used to Calculate Mediated Effects of Program Participation

Outcome	A Path Composite Discussion Quality Rating	C Prime Path Academic Vocabulary (Post)	C Path Academic Vocabulary (Post)
Treatment	0.393*	1.521**	1.180*
	(0.171)	(0.506)	(0.549)
Academic vocabulary (pre)	0.138***	0.700***	0.696***
	(0.032)	(0.024)	(0.024)
Composite discussion score			0.516
			(0.368)
Intercept	−2.809***	5.928***	6.206***
	(0.576)	(0.556)	(0.585)
Level 2 variance (teaching team)		1.804	1.675
		(0.324)	(0.311)
Residual		28.27***	28.28***
		(0.536)	(0.537)
N	1,442	1,442	1,442

Note. Multilevel mixed-effects linear regression was used to fit c path and c prime path models (STATA *xtmixed* command); between-effects linear models were used to estimate a path. Standard errors in parentheses.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 8
Estimates of Confidence Intervals for Indirect Effects, Direct Effects, and Total Effects of Treatment on Academic Vocabulary Mediated by Improved Discussion Generated by Bootstrap Resampling With 500 Replications

	Coefficient	Standard Error	*z*	*P* Value	95% Confidence Interval	
Indirect effect	0.20	0.09	2.18	.03	0.03	0.38
Direct effect	1.18	0.30	3.89	>.001	0.59	1.78
Total effect	1.39	0.29	4.76	>.001	0.81	1.96

This math lesson fell on Day 2 of a unit focused on single-gender education. On Day 1, the language arts teacher had guided students through an initial reading and discussion of the informational text, *Single Gender Education: Are Academics More Important Than Social Learning?*, and introduced students to meanings of the following target words: *paradigm*, *comprise, gender*, *conduct*, *adapt*.

Word Generation Randomized Trial

The Word Generation math lesson required mathematical thinking about problems situated in the context of the debate over single gender education. Students were presented with two questions requiring basic calculations based on an understanding of proportion, multiplication, and division. To interpret the questions, students had to understand the meaning of three of the target academic vocabulary words in the context of the problems. For example, one question read:

> The year is 2015. In the U.S., 90 schools have adopted the single-**gender paradigm**. All-girls schools **comprise** one-third of those 90 schools. How many all-girls schools are there? Show or explain how you got your answer.

To understand the question, students had to understand the meaning of single-*gender* and they had to integrate the meanings of *paradigm* and *comprise* with the context of the question. The teacher facilitated a discussion that moved back and forth between clarifying the math with discussion about the significance of the questions about single-gender education. After talking students through procedures for finding the answer (clarification of denominator, multiple choice strategy of eliminating certain choices, selecting the best possible answer), the teacher concluded with an open-ended discussion question: "The number of single-gender public schools increased dramatically between 1995 and 2008. How do you explain this dramatic increase?"

This discussion received high scores for all dimensions of classroom discussion, including the highest score for the dimensions support for participation (3) and student engagement (3). For teacher talk moves and substantive contributions, the discussion received scores of 5 and 4, respectively (range, 1-5). While the teacher posed authentic, open questions and pressed students to articulate their reasoning and students' contributions included substantive explanations of their thinking, students did not relate their contributions to those of their peers.

Three observations about this discussion are of note. First, we can see how the nature of the exchanges interacted with the content to support student learning of target words. Second, student engagement was high, as students were interested not just in solving the problem but also in understanding the impact that single-gender education might have on their lives. For example, while talking through how to solve the problem, one student interjected, "I'm just curious. Is that what's going to happen in a few years?" While this widespread engagement can be explained in part by the particular dilemma under discussion, it also seemed to reflect the teacher's consistent effort to press students to explain or defend their reasoning. For example:

> Teacher: So do we need schools that are single gender?
> Student 1: If a school has just boys, it will help focus on what boys need. If it's just girls it will focus on what girls need.

Teacher: So are some of those needs academic? Should they be different?
Student 1: I'm not sure what you mean.
Teacher: Should they be different academically? Should what boys learn in their classes be different from what girls learn in their classes?
Student 2: Well, sometimes boys need a little extra attention.
Teacher: What kind of attention?

By following up students' contributions with questions asking students to clarify or think more deeply about the issue, the teacher promoted a high degree of engagement. Thus, the math problem dealt with an issue that is of importance to students and the teacher facilitated a discussion that promoted their involvement and interaction. It seems likely that these students would not only experience more encounters with target words (especially *gender* and *comprise*) but that these encounters would effectively build strong representations of word meanings because students would attend closely to their meanings and uses.

Second, it is interesting to note that the teacher's language during this high-quality discussion included use of other general academic words such as *eliminate* and *academic*. Thus, in some cases enriched discussion opens up opportunities to hear other academic words, creating more encounters with these words over time and in a greater variety of contexts.

Example 2:
Topic: Should owning handguns be legal? Unit 20
Social Studies
Words: *scheme*, *subsequently*, *dominant*, *import*, *commission*

This social studies lesson fell on Day 4 of a unit focused on whether or how it should be legal to own a handgun. On Day 1, the language arts teacher had guided students through an initial reading and discussion of an informational text about the shootings at Virginia Tech and had introduced students to meanings of the five target words.

During the discussion, the teacher used several of the target words effectively, creating additional encounters likely to enrich students' representations of those words. For example, in relation to one of the debate positions, she explained, "You can import guns. Remember importing is bringing in from another country, guys. A lot of times guns that are *imported* aren't even legal." In response to a student who argued that guns are necessary for self-defense, the teacher responded "It could be self-defense in the case of *domestic* abuse" and later conceded "this is something to keep in mind. Maybe they don't need to be illegal but there should be more *monitoring*."

Note that as the teacher facilitated the discussion, she not only supported student learning of target words, for example, by using *import*, but also created encounters with other academic words such as *domestic* and *monitoring*. More examples of transcribed classroom discussion (with video) from Word Generation classes in each content area are available at www.wordgeneration.org (although these were not classrooms that participated in the study reported here).

Discussion

The most provocative finding from this study is that teachers implementing the Word Generation program have dramatically higher average classroom discussion ratings than control teachers. Interpreting this finding requires care since we know that classroom discussion varies across the school year within teachers' classrooms as well as across classroom settings. Our classroom observations were conducted at times when control teachers indicated that they would be engaged in instructional activities that would generate extended discussion. The classroom discussion found in these control classes was often brief and poor but not at all out of line with the levels of discussion reported in other large-scale studies (Applebee et al., 2003; Nystrand & Gamoran, 1991). Discussion in Word Generation classrooms was rated dramatically higher. Word Generations teachers had participated in relatively limited professional development that highlighted the importance of discussion, and students had access to curricular materials related to social issues that were designed to elicit student opinion; both factors undoubtedly contributed to more lively, engaged, and talkative classrooms. These data do not tell us anything about classroom discussion in the treatment schools during the rest of the day. We doubt that levels of discussion were as high when treatment teachers were not implementing the curriculum since Word Generation materials are intended to generate engaged discussion, though we might hope that the impact of the professional development and experience using the program spilled over into other classes to some extent. More needs to be done to understand the relative impact of the professional development and the curricular materials in influencing classroom discussion both during Word Generation implementation and during other classroom periods in the treatment schools.

A closer look at the classroom observation data demonstrates that the biggest contrasts between quality ratings in treatment and control schools occurred at the lower level of the quality continuum. Figure 1 illustrates the difference in the frequency with which treatment and control classrooms received the lowest scores on Composite Discussion Quality Ratings but a much less pronounced difference in frequency scores at higher levels. These data suggest engaging materials and a light professional development program may significantly improve classroom discussion in classes where

discussion is of poor quality. In our ongoing work, we are exploring the kinds of coaching and professional development teachers need to use these materials most effectively and how they might transfer skills they develop while implementing the Word Generation program across instructional contexts.

Despite strong treatment effects on student participation in discussion, the program only had a small effect on student knowledge of targeted academic vocabulary and no effect on the Gates-MacGinitie Vocabulary measure. Estimated treatment effect in our HLM models showed that students in the treatment schools scored roughly 1.5 points higher on the curriculum-based posttest controlling for pretest and a wide range of covariates. Given that these calculations are only based on 36 out the 120 words that were taught (30%), we can extrapolate this to mean that students in the treatment schools learned an average of 5 words more than the comparison students during the year as measured by the synonym identification assessment. One reason for these disappointing results might be that multiple choice synonym assessments fail to assess many dimensions of word knowledge (Pearson, Heibert, & Kamil, 2007). Our ongoing analysis with a broader range of assessment types suggests that estimated treatment effects will vary depending on the types of assessments used (Lawrence, Pare-Blagoev, Lawless, Deane, & Li, 2013). That being said, the results reported here replicate a general finding that standardized measures of general vocabulary knowledge rarely show effects from targeted vocabulary interventions (Elleman, Lindo, Morphy, & Compton, 2009), leading to the recommendation to include curriculum-based measures in order to detect intervention effects (NICHD, 2000). Elleman and colleagues (2009) found that the mean effect size of interventions on learning of words explicitly taught was d = .79 and that discussion quality was related to effect size.

Clearly, the largest short-term vocabulary gains are to be expected from curricula that focus exclusively on vocabulary and from those that teach many words (Biemiller & Boote, 2006). Vocabulary interventions that teach 10 to 15 words per week can be expected to produce larger effects than programs like Word Generation, which taught only a few words per week and focused more centrally on introducing new pedagogical practices into the teacher's repertoire. Thus, the finding that the program produced larger impacts on quality of discussion than on word learning is not surprising. In current work with an enriched version of Word Generation we are exploring a greater array of possible student-level effects, including academic language, perspective taking, reasoning, argumentation, and reading comprehension; we are also studying in greater detail the impact on teachers' instructional approaches, both during Word Generation lessons and at other times.

One of the most interesting results from this study is that improved discussion mediated the treatment effect on student word learning. We do not know of any prior empirical study that has established that improved

discussion mediates treatment effects on student word learning from a vocabulary intervention. On the other hand, academic discussion provides precisely the contexts that are known to support vocabulary learning, as exemplified in the extracts of classroom discussion presented previously.

This study represents a first attempt to influence the nature of classroom discourse by introducing discussable topics in addition to promoting more discussion-based pedagogies. The choice of dilemmas as the organizing themes for each week of the curriculum serendipitously created one of the conditions that foster authentic discussion, namely, the absence of an easy or known answer to the question. In responding to these dilemmas, teacher views were not privileged over student views, and thus the participation structure could be authentically egalitarian, leading to many student-student as well as teacher-student exchanges.

Furthermore, the study was carried out in urban schools that faced all the challenges typical of such schools: inadequate resources, entrenched teachers, demotivated students, and variable levels of instructional leadership. That reliable changes could be effected under these circumstances suggests high external validity for the findings. Of course we recognize that the effects obtained were small and even if multiplied over several years, would be inadequate to close the achievement gap between these low SES students and their middle-class peers. Nonetheless, we hope we have illuminated one mechanism that can contribute to broader changes in schooling effectiveness for students most at risk of academic failure.

There are several limitations to this study. The number of students who contributed both pre- and posttest data was negatively affected by scheduling problems and transiency in these urban schools. We did not receive student-level data from one of the districts and thus could not control for individual demographic variables in our models. We did not collect baseline information about teachers' experience with using discussion techniques or implementing a shared curriculum. There are clearly mechanisms beyond discussion that might have affected student outcomes that we cannot account for. We are not sure to what extent teachers in treatment schools used Word Generation approaches in other classes; more should be done to understand teachers' perspectives on this work and what they learned from using the Word Generation program. We did not have enough classroom observations of each team to create reliable estimates of discussion quality for each teaching team; we conducted our mediation analysis at the school level. Nonetheless, we have shown that it is possible to increase amount and quality of discussion, even in classrooms in challenging urban districts, and that doing so is related to increases in student learning of targeted vocabulary.

Notes

This work was supported by grant number R305A090555, Word Generation: An Efficacy Trial from the Institute of Educational Sciences, US Department of Education (Catherine Snow, PI).

[1]We replicated the basic components of the analyses reported here with multiply imputed data sets; pre- to post-changes and treatment coefficients (using the *mi estimate* command in Stata with imputed data set) were similar to those reported here. We were not confident in our ability to impute the multilevel data structure correctly and then conduct multilevel models and mediation analysis with the multiply imputed multilevel data, so we did not. However, the descriptive results using the multiply imputed data sets confirm the findings presented here and gave us no reason to think that missing data influenced our results.

[2]Since the program is implemented by groups of teachers who teach the same sets of students across content areas, we requested course schedules from all schools to understand the way schools organized teachers across content areas to deliver instruction. We found that for the most part, cross-content teaching teams were identical to grade-level teams within each building.

[3]The correlation between the weighted and nonweighted school-level discussion scores is $r = 0.98$. The weighted score is a better metric as it is not influence as much by difference in the number of content area classes that were observed in each school, which was not balanced due to schedule and differences in school sizes.

[4]Vocabulary data were collected at the individual level, but we were interested in understanding the treatment effect at the group level. In order to account appropriately for the nested structure of the data (one, two, or three grades within schools within districts), we used multilevel modeling techniques to estimate the intent-to-treat effect size (Raudenbush & Bryk, 2002). We were particularly interested in understanding which models would be most appropriate for the analysis, given the range of intensity of implementation across grade-level teaching teams. We found that the model with school and grade level variance specifications was only marginally better (–2 log likelihood [LL] = 9,648.20) than the same model using only teaching team as the variance component (–2 LL = 9,654.23) and was not warranted given the necessary addition of two variance parameters (i.e., the variance parameter associated with schools and the covariance parameter; $\boldsymbol{\Delta}$ –2LL = 6.03; $df = 2$, $p = ns$).

[5]$\Delta = \delta_T - \delta_C = \frac{(\mu_{T,post} - \mu_{T,pre}) - (\mu_{C,post} - \mu_{C,pre})}{\sigma}$.

[6]We calculated pooled standard deviation on Table 5 with the following equation:

$$S_{pooled} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_1^2}{(n_1 + n_2 - 2)}}$$

References

Alvermann, D. E., & Hayes, D. A. (1989). Classroom discussion of content area reading assignments: An intervention study. *Reading Research Quarterly*, *24*, 305–335.

Applebee, A., Langer, J., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal*, *40*, 685–730.

Beck, I., McKeown, M., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd ed.). New York, NY: Guilford Press.

Biemiller, A., & Boote, C. (2006). An effective method for building meaning vocabulary in primary grades. *Journal of Educational Psychology*, *98*(1), 44–62.

Bloom, H., Richburg-Hayes, L., & Black, A. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*, 30.

Bolger, D., Balass, M., Landen, E., & Perfetti, C. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, *45*, 122–159.

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

Chapin, S., O'Connor, M., & Anderson, N. (2003). *Classroom disucssions: Using math talk to help students learn: Grades 1-6*. Sausalito, CA: Math Solutions Publications.

Common Core State Standards Initiative. (2010). *Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects*. Washington, DC: National Governors Association Center for Best Practices and the Council of Chief State School Officers.

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

Duke, N. K., Pearson, P. D., Strachan, S. L., & Billman, A. K. (2011). Essential elements of fostering and teaching reading comprehension. In S. J. Samuels & A. E. Farstrup (Eds.), *What research has to say about reading instruction* (pp. 51–93). Newark, DE: IRA.

Echevarria, J., Short, D., & Powers, K. (2006). School reform and standards-based education: A model for English-language learners. *The Journal of Educational Research*, *99*, 195–211.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall.

Elizabeth, T., Ross, T., Snow, E., & Selman, R.L. (2012). Academic discussions: An analysis of instructional discourse and an argument for an integrative assessment framework. *American Educational Research Journal*, *49*(6), 1214–1250. doi:10.3102/0002831212456066

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

Elmore, R. F., Forman, M. L., Stosich, E. L., & Bocala, C. (2013). *The internal coherence assessment protocol & developmental framework: Building the organizational capacity for instructional improvement in schools*. Cambridge, MA: Harvard University.

Freebody, P., & Anderson, R. C. (1983). Effects on text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Reading Behavior*, *15*, 19–39.

Gamoran, A., & Nystrand, M. (1991). Background and instructional effects on achievement in eighth-grade English and social studies. *Journal of Research on Adolescence*, *1*, 277–300.

Graves, M. F. (2000). A vocabulary program to complement and bolster a middle-grade comprehension program. In B. M. Taylor, M. F. Graves, & P. Van den Broek (Eds.), *Reading for meaning: Fostering comprehension in the middle grades* (pp. 116–135). Newark, DE: International Reading Association.

Hiebert, J., & Wearne, D. (1993). Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal*, *30*, 393–425.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, *5*(5), 602–619.

Krull, J., & MacKinnon, D. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*(2), 249–277.

Larson, B. E. (2000). Classroom discussion: A method of instruction and a curriculum outcome. *Teaching and Teacher Education*, *16*(5), 661–677.

Lawrence, J. F., Pare-Blagoev, E., Lawless, R. R., Deane, P., & Li, C. (2013). *Contributions of depth measures to an assessment of the Word Generation vocabulary intervention*. Paper presented at the Annual meeting of the American Educational Research Association, San Francisco, CA.

Lawrence, J. F., Capotosto, L., Branum-Martin, L., White, C., & Snow, C. E. (2012). Language proficiency, home-language status, and English vocabulary development: A longitudinal follow-up of the Word Generation program. *Bilingualism: Language and Cognition*, *15*(3), 437–451.

Lemke, J. L. (1990). *Talking science: Language, learning, and values* (Vol. *1*). Norwood, CT: Ablex Publishing Corporation.

Lesaux, N. K., Kieffer, M. J., Faller, S. E., & Kelley, J. G. (2010). The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools. *Reading Research Quarterly*, *45*, 196–228.

Lin, T.-J., Anderson, R. C., Hummel, J. E, Jadallah, M., Miller, B. W., Nguyen-Jahiel, K., . . . Dong, T. (2012). Children's use of analogy during collaborative reasoning. *Child Development*, *83*, 1429–1443.

Lipman, M. (1976). Philosophy for children. *Metaphilosophy*, *7*, 17–33.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie reading test technical report: Forms S and T*. Chicago, IL: The Riverside Publishing Company.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*, 41–62.

Mancilla-Martinez, J., & Lesaux, N.K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology*, *102*, 701–711.

Matsumura, L. C., Slater, S. C., & Crosson, A. (2008). Classroom climate, rigorous instruction and curriculum, and students' interactions in urban middle schools. *The Elementary School Journal*, *108*, 293–312.

McKeown, M. (1985). The acquisition of word meaning from context by children of high and low ability. *Reading Research Quarterly*, *20*, 482–496.

Mercer, N., Wegerif, R., & Dawes, L. (1999). Children's talk and the development of reasoning in the classroom. *British Educational Research Journal*, *25*(1), 95–111.

Michaels, S., O'Connor, C., & Resnick, L. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education*, *27*, 283–297.

Murphy, P., Wilkinson, I., Soter, A., Hennessey, M., & Alexander, J. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology*, *101*, 740–764.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: U.S. Government Printing Office.

Nystrand, M., & Gamoran, A. (1991). Instructional discourse, student engagement, and literature achievement. *Research on the Teaching of English*, *25*, 261–290.

Nystrand, M., Wu, L., Gamoran, A., Zeiser, S., & Long, D. (2003). Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse processes*, *35*, 135–198.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring strategies. *Cognition & Instruction*, *1*, 117–175.

Pearson, P., Hiebert, E., & Kamil, M. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly*, *42*(2), 282–296.

Preacher, K. J., & Hayes, A. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*, 879–891.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. *1*). Thousand Oaks, CA: Sage Publications, Inc.

Reisman, A. (2012). Reading like a historian: A document-based history curriculum intervention in urban high schools. *Cognition and Instruction*, *30*, 86–112.

Reznitskaya, A., Anderson, R., McNurlen, B, Nguyen-Jahiel, K., Archodidou, A., & Kim, S. (2001). Influence of oral discussion on written argument. *Discourse Processes*, *32*, 15–175.

Reznitskaya, A., Kuo, L., Clark, A., Miller, B., Jadallah, M., Anderson, R. C., & Nguyen-Jahiel, K. (2009). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education*, *39*, 29–48. doi:10.1080/0305764 0802701952

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*, 325–344.

Snow, C. E., Porche, M. V., Tabors, P. O., & Harris, S. R. (2007). *Is literacy enough? Pathways to academic success for adolescents*. Baltimore, MD: Paul H. Brookes, Publishing Co.

Soter, A. O., Wilkinson, I. A., Murphy, P. K., Rudge, L., Reninger, K., & Edwards, M. (2008). What the discourse tells us: Talk and indicators of high-level comprehension. *International Journal of Educational Research*, *47*, 372–391.

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

Wegerif, R., Mercer, N., & Dawes, L. (1999). From social interaction to individual reasoning: an empirical investigation of a possible socio-cultural model of cognitive development. *Learning and Instruction*, *9*, 493–516.

What Works Clearinghouse. (2008). *Evidence standards for reviewing studies* (Technical Report). Washington, DC: Institute for Educational Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_version1_standards .pdf.

Wolf, M. K., Crosson, A. C., & Resnick, L. B. (2005). *Accountable talk in reading comprehension instruction* (CSE Technical Report 670). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning and Technology*, *13*, 49–67.

Manuscript received October 4, 2013
Final revision received August 7, 2014
Accepted January 14, 2015