Impact of a Large-Scale Science Intervention Focused on English Language Learners

Article *in* American Educational Research Journal · March 2016

DOI: 10.3102/0002831216637348

CITATIONS
24

READS
395

7 authors, including:

Lorena Llosa
New York University
32 PUBLICATIONS **253** CITATIONS
SEE PROFILE

Feng Jiang
University of Arkansas
10 PUBLICATIONS **82** CITATIONS
SEE PROFILE

Christopher D. Van Booven
University of Massachusetts Dartmouth
5 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Michael J Kieffer
New York University
52 PUBLICATIONS **2,135** CITATIONS
SEE PROFILE

American Educational Research Journal
April 2016, Vol. 53, No. 2, pp. 395–424
DOI: 10.3102/0002831216637348

Impact of a Large-Scale Science Intervention Focused on English Language Learners

Lorena Llosa
Okhee Lee
New York University
Feng Jiang
University of Arkansas
Alison Haas
Corey O'Connor
Christopher D. Van Booven
Michael J. Kieffer
New York University

The authors evaluated the effects of P-SELL, a science curricular and professional development intervention for fifth-grade students with a focus on English language learners (ELLs). Using a randomized controlled trial design with 33 treatment and 33 control schools across three school districts in one state, we found significant and meaningfully sized intervention effects on a researcher-developed science assessment and the state science assessment. Subgroup analyses revealed that the P-SELL intervention had a positive and significant effect for each language proficiency group (ELLs, recently reclassified ELLs, former ELLs, and non-ELLs) on the researcher-developed assessment. The intervention also had a positive effect for former ELLs and non-ELLs on the state science assessment, but for ELLs and recently reclassified ELLs, the effect was not statistically significant.

KEYWORDS: science achievement, English language learners, randomized controlled trial, inquiry-based science, elementary school science

The imperative that all students, especially English language learners (ELLs), achieve high academic standards in science is becoming ever more urgent and complex as a result of three key factors: the growing diversity of the U.S. student population, persistent science achievement gaps, and the increasing demands of high-stakes assessment and accountability in science. First, while student diversity has been increasing steadily, ELLs make up the fastest growing student population in the United States. According to the 2010 U.S. Census (U.S. Census Bureau, 2012), 21% of children 5 to

17 years old spoke a language other than English at home. During the 2011–2012 school year, students with limited English proficiency (the term used by the federal government), or ELLs, constituted 9% of public school students (National Center for Education Statistics [NCES], 2014). Despite rapidly growing student diversity, few teachers report feeling prepared to provide science instruction for diverse student groups, including ELLs (Banilower et al., 2013). Second, science achievement gaps persist among demographic subgroups. On the National Assessment of Educational Progress (NAEP) between 1996 and 2011, science achievement gaps between ELLs and non-ELLs have remained largely consistent and wide (NCES, 2012). Thus, developing interventions and preparing teachers to meet the academic needs of ELLs in science is a major concern of U.S. education reform in the 21st century. Third, there is a growing role of science in accountability systems at federal and state levels. Since the 2007–2008 school year, No Child Left Behind requires that each state administer science assessments at least one time during Grades 3 to 5, Grades 6 to 9, and Grades 10 to 12. In some states, such as the state where this study took place, science counts toward a school's annual evaluation.

Lorena Llosa is an associate professor in the Steinhardt School of Culture, Education, and Human Development at New York University, 239 Greene Street, 514, New York City, NY 10003, USA; e-mail: *lorena.llosa@nyu.edu*. Her research aims to improve the educational opportunities of English language learners in schools. Her interests include language assessment, language and content integration, and program evaluation.

Okhee Lee is a professor in the Steinhardt School of Culture, Education, and Human Development at New York University. Her research areas include science education, language and culture, and teacher education.

Feng Jiang is a research associate in the Office of Innovation for Education at the University of Arkansas. His interests include science education, nature of science, and quantitative analysis of education data.

Alison Haas is a research associate and doctoral student in the Steinhardt School of Culture, Education, and Human Development at New York University. Her research interests focus on elementary education including curriculum development, teacher professional development, educational policy, and educational technology.

Corey O'Connor is the project manager of the P-SELL project in the Steinhardt School of Culture, Education, and Human Development at New York University. His interests include life sciences and science education.

Christopher D. Van Booven is a PhD candidate in the Steinhardt School of Culture, Education, and Human Development at New York University. His research focuses on language and science learning practices in out-of-school contexts and how these practices relate to learning and instruction in schools.

Michael J. Kieffer is an associate professor of literacy education in the Steinhardt School of Culture, Education, and Human Development at New York University. His research focuses on the language and literacy development of students from linguistically diverse backgrounds.

The convergence of these key factors highlights that educational interventions to promote science learning for all students, including ELLs, are greatly needed in the context of high-stakes science assessment and accountability policy. However, a recent research synthesis of elementary science programs found very few studies that both met inclusion criteria for rigorous designs and showed a positive impact for improving student learning (Slavin, Lake, Hanley, & Thurston, 2014). In addition, none of the studies included in the synthesis addressed the ELL population.

This study aims to address this gap in the literature by examining the impact of P-SELL (Promoting Science Among English Language Learners; Lee & Llosa, 2011–2015), a curricular and professional development intervention aimed at improving science achievement of all students and ELLs in particular. The P-SELL intervention consists of a standalone, year-long science curriculum that addresses state science standards for fifth-grade students and teachers as well as professional development workshops for teachers focusing on curriculum implementation. The curriculum is designed to promote students' scientific inquiry and understanding while providing language development strategies. The study took place in three demographically and geographically disparate school districts in a state in which fifth-grade science is tested and counts toward school accountability. This study was motivated by the urgent need for ELLs to have access to equitable learning opportunities so that they can be successful in school and be ready for college and careers.

Literature Review

ELLs in schools need to develop their English proficiency in order to benefit from content area instruction in English. At the same time, content area instruction provides ELLs with a context for language learning and an authentic purpose for communication. The fact that content area learning in general and science learning in particular provide a meaningful context for ELLs' language development has long been recognized (Lee & Fradd, 1998; Rosebery, Warren, & Conant, 1992). Much of the early literature on effective science instruction for ELLs focused primarily on hands-on activities and their potential to make science concrete and experiential. The focus later shifted to the need to integrate cognitively challenging science inquiry practices with an explicit attention to language and literacy strategies, including the use of ELLs' home language and culture as instructional supports (Buxton & Lee, 2014; Fathman & Crowther, 2006; Janzen, 2008; Lee, 2005; Rosebery & Warren, 2008). Since the development of the Next Generation Science Standards and the Common Core State Standards, researchers have argued for a deeper integration of science and language learning that goes beyond the incorporation of hands-on activities and language strategies and instead focuses on the need for all students including ELLs to use

language while engaging in science and engineering practices (Lee, Quinn, & Valdés, 2013).

Informed by the early literature on integration of science and language instruction, interventions have been developed over the past couple of decades to address the needs of ELLs. However, the majority of studies available about interventions for upper elementary and middle grade ELLs have been descriptive in nature, with few studies testing the impact of interventions. Those studies that aimed to evaluate the impact of interventions often had methodological limitations such as the lack of a control group (e.g., Lee, Deaktor, Enders, & Lambert, 2008; Santau, Maerten-Rivera, & Huggins, 2011). Thus, empirical evidence of the beneficial impact of science and language integration for ELLs was limited.

In more recent years, experimental and quasi-experimental studies have been conducted to examine the effectiveness of interventions focused on science and language development for ELLs. Some interventions have focused primarily on ELLs' language development in the context of science learning (Echevarria, Richards-Tutor, Canges, & Francis, 2011; Zwiep & Straits, 2013), others have focused on both language development and science learning (August, Branum-Martin, Hagan, & Francis, 2009; August et al., 2014; Lara-Alecio et al., 2012), and still others have focused primarily on science learning while attending to language development (Maerten-Rivera, Ahn, Lanier, Diaz, & Lee, in press). These studies are summarized next.

Echevarria et al. (2011) examined the efficacy of a model of instruction, the Sheltered Instruction Observation Protocol (SIOP) Model, in middle school science. The goal of the intervention was to provide teachers with strategies to make science content comprehensible to ELLs and develop academic language. A small, cluster randomized trial with randomization at the school level was attempted. Initially 10 middle schools in one large urban district were randomly assigned to either the treatment (SIOP Model) or a control, but two schools in the control condition dropped out of the study. In the end, the study involved 8 schools (5 SIOP and 3 comparison), 12 teachers (8 SIOP and 4 comparison), and 1,021 students (649 SIOP and 372 comparison). Teachers in the SIOP condition received training and then taught four science units over eight weeks using lesson plans and teaching methods that followed the SIOP Model. Multilevel analyses revealed that the intervention had no impact on students' academic language or reading comprehension of passages about science topics. The authors propose that the lack of impact might have been due to the short duration of the intervention, the limited opportunities for professional development, and the varied levels of cooperation, interest in the study, and implementation across the eight SIOP teachers.

Zwiep and Straits (2013) developed and implemented a blended inquiry science and English Language Development (ELD) program in a large urban school district. This program was used during ELD time as an alternative to

traditional approaches to language teaching in a district that did not support science instruction in elementary schools. In this blended program, teachers used science as the context for language development and provided ELLs with opportunities to develop English proficiency through participation in inquiry-based science (for details of the program, see Zwiep, Straits, Stone, Beltran, & Furtado, 2011). This quasi-experimental study involved three elementary schools with 60 teachers and over 2,000 K–5 students in the treatment group. These students' gains in English and science achievement were compared to those of students at two comparison schools that were using the district's ELD curriculum. Over the four years of the project, results indicated modest but statistically significant improvement on state-mandated English language proficiency (ELP), English language arts (ELA), and science assessments for students who participated in the blended program.

August et al. (2009) designed and tested the effectiveness of Quality English and Science Teaching (QuEST), a program designed to simultaneously support the science learning and academic language development of middle grade ELLs. Building on science textbooks and workbooks that were already in use in middle schools, the QuEST curriculum involved the infusion of the 5E science inquiry model and the direct instruction of both general and topic-specific academic vocabulary as well as linguistic scaffolding for ELLs (e.g., enhanced use of visual materials and graphic organizers, previews of science experiments for ELLs, and use of instructional conversations during science investigations). The curriculum consisted of two units on living systems and the environment that lasted nine weeks. Using a cluster randomized trial, the study involved 10 sixth-grade science teachers in five middle schools in a large school district. For each teacher, two sections were randomly assigned to the treatment group that used the QuEST curriculum, and two sections were randomly assigned to the control group that used the district curriculum. Teachers participated in workshops and weekly mentoring on the QuEST curriculum. Posttest differences favoring the treatment group sections were statistically significant for the researcher-developed measures of vocabulary and science. When they examined the effect of the treatment on ELLs only, they also found significant differences favoring the treatment group for both vocabulary and science.

As a follow-up study, August et al. (2014) tested the effectiveness of QuEST 2, designed to help ELLs and their English-proficient classmates develop academic language in science, as required by the Common Core State Standards. Using a cluster randomized trial, 60 sections taught by 15 teachers in seven middle schools were randomized within teacher to the treatment or the control condition. Both treatment and control sections used the same district-adopted textbook, workbook, and labs. The treatment sections used two additional components: a curriculum consisting of inquiry-based lessons explicitly addressing academic language and professional development for teachers. The 15-week intervention had a positive

impact on a researcher-developed measure of academic language, but the treatment effect for the researcher-developed science measure was not statistically significant. Similarly, for ELLs, a small group difference between treatment and control sections was found for academic language but not for science.

Lara-Alecio et al. (2012) also studied the effects of an intervention focused on both inquiry-based science and language development. Unlike August et al. (2009, 2014), their intervention comprised a year-long curriculum and used state and district assessments as their outcome measures. Their fifth-grade intervention consisted of ongoing professional development and specific instructional science lessons with inquiry-based learning, direct and explicit vocabulary instruction, integration of reading and writing, and enrichment components including integration of technology, take-home science activities, and mentoring by university scientists. Two schools were randomly selected from 10 schools in the school district and then randomly assigned to the treatment condition (with 166 students) and control condition (with 80 students). Because there was a low response rate for participation among control teachers, additional teachers had to be invited into the control group, and thus their study ended up being quasi-experimental. Results suggested a significant and positive intervention effect in favor of the treatment students as reflected in higher performance in district-wide benchmark tests of science and reading and a measure of oral reading fluency. However, there was no significant difference between the treatment and control groups on the state science assessment.

Maerten-Rivera et al. (in press) also examined the impact of a year-long science intervention for fifth-grade students. However, this intervention was implemented at a much larger scale than that in Lara-Alecio et al. (2012)'s study and focused primarily on students' science achievement. The intervention, a prior iteration of the intervention in the current study, addressed science inquiry and language development strategies for ELLs. The study involved a randomized controlled trial with 64 randomly selected schools in one school district in a Southeastern state: 32 schools randomly assigned to the treatment group and 32 schools randomly assigned to the control group. After the first year of implementation, 1 treatment school withdrew. Thus, a total of 31 treatment schools and 32 control schools participated over the three-year implementation. The study involved all of the fifth-grade teachers (about 350 over three years) and students (about 6,000 each year) from the 63 schools. No significant difference was found between the treatment and control groups on the state science assessment in Year 1 of the study, but there was a significant difference across the two groups in Years 2 and 3, with the treatment group outperforming the control group. The difference between the two groups widened from Year 2 to Year 3. The percentage of students classified as proficient in science according to the state science assessment was also examined according to students'

English proficiency classification: ELLs, former ELLs, and non-ELLs. Similar to the overall achievement results, the differences increased each year, with the greatest differences in Year 3. However, the interaction between treatment and language classification was not statistically significant, indicating the effect of the intervention did not vary based on students' language classification.

In summary, several studies have been conducted in recent years that have employed an experimental or quasi-experimental design to investigate the impact of interventions integrating science and language instruction focused on ELLs. Overall, results have been promising but inconclusive, with many of the studies being more successful at impacting language development than science achievement. Methodologically, except for Maerten-Rivera et al. (in press), all of these studies were relatively small, involving only a few schools (ranging from 2 to 8). In addition, many of the studies used researcher-developed measures only. Even though all of these studies, except for Zwiep and Straits (2013), implemented their interventions with students ranging from English proficient to ELLs, only Echeverria et al. (2011) and Maerten-Rivera et al. (in press) examined the impact of the interventions for ELLs, former ELLs, and non-ELLs. Saunders and Marcelletti (2013) emphasize that former ELLs "should be factored into study designs, data collections, and explanations of results" (p. 154). They argue that not including former ELLs "runs the risk of inadvertently misrepresenting the potential of ELs" (p. 154). Finally, only August et al. (2009, 2014) and Maerten-Rivera et al. employed an experimental design with random assignment to condition, allowing for a direct attribution of the treatment effect to the intervention.

The current study addresses many of the limitations of prior work. It employed a randomized controlled trial design to investigate the impact of P-SELL, a year-long, standalone fifth-grade science curriculum (a revised version of the curriculum used in Maerten-Rivera et al., in press). The study was implemented large-scale, with 66 randomly selected schools (33 treatment and 33 control) across three school districts in one state. The study used both the high-stakes state science assessment as an outcome measure and a researcher-developed science assessment that was administered pre and post and allowed for a pre-measure of science achievement. The study examined the effect of the intervention on science achievement for all students and for students of varying levels of English proficiency (ELL, recently reclassified ELL, former ELL, and non-ELL). Specifically, this study addressed the following research questions:

What was the effect of the P-SELL intervention on fifth-grade students' science achievement compared to "business as usual"?

Research Question 1: Was the P-SELL intervention beneficial, on average, for students?

Research Question 2: Was the P-SELL intervention beneficial for ELLs, recently reclassified ELLs, former ELLs, and non-ELLs?

The study reported here presents the results of the first year of implementation of the three-year intervention. By addressing many of the limitations of prior research, the current study is uniquely positioned to yield important information about the impact of large-scale interventions aimed at improving science achievement of all students, including ELLs, recently reclassified ELLs, former ELLs, and non-ELLs, in elementary schools.

The P-SELL Intervention

Our curricular and professional development intervention is conceptually grounded in the literature on effective science instruction with ELLs (Buxton & Lee, 2014; Fathman & Crowther, 2006; Janzen, 2008; Lee, 2005; Rosebery & Warren, 2008). The literature on educative curriculum materials (Davis et al., 2014; Davis & Krajcik, 2005; Drake, Land, & Tyminski, 2014) and effective professional development (Desimone, 2009; Garet, Porter, Desimone, Birman, & Yoon, 2001) guided the implementation of the intervention. The curriculum itself and its implementation were also informed by our previous research and experience with an earlier iteration of the curriculum (Maerten-Rivera et al., in press). Next, we describe the student and teacher components of our intervention.

Student Components of the P-SELL Intervention

The P-SELL intervention consists of a comprehensive, standalone, yearlong fifth-grade science curriculum. The curriculum includes (a) consumable student books, (b) science supplies, and (c) supplementary materials on the project website. The curriculum adopts a standards-based and inquiry-oriented approach for all students, especially ELLs, with a focus on three key features.

First, the P-SELL curriculum's standards-based approach aligns with state science standards and high-stakes science assessment administered at fifth grade. In the state where the study took place, the state standards consist of 18 "big ideas" according to four "bodies of knowledge," including the nature of science, earth and space science, life science, and physical science. These standards are assessed using the state science assessment, and student performance on this assessment is part of the state's accountability system. Our curriculum is organized around these big ideas. Each chapter begins by identifying the science content standards and benchmarks addressed. Furthermore, each hands-on inquiry activity, reading passage, and writing section designates the science content standard(s) and benchmark(s) addressed.

Second, the P-SELL curriculum is based on an inquiry-oriented approach. Science inquiry serves both as a goal of science learning and as a means through which students develop scientific understanding of the big ideas in the state science standards. The student book is designed to follow a progression from teacher-directed instruction to student-initiated

inquiry. That is, by providing more structure in earlier chapters and a more open-ended approach in later chapters, the curriculum fosters student initiative and exploration. In addition, as they complete inquiry activities, students are encouraged to design their own extension inquiry activities and apply key science concepts to everyday events or phenomena in home and community contexts.

Finally, the P-SELL curriculum addresses the learning needs of ELLs by providing guidance and scaffolding for English language development. Each chapter starts with key science terms in the three primary languages spoken by students of the participating school districts: English, Spanish, and Haitian Creole. The chapter then introduces key science concepts by connecting them with students' prior knowledge or experiences in their home and community contexts as well as their knowledge from previous chapters. The curriculum uses multiple modes of representation in textual and graphic formats (e.g., students write extensively in the student book) and oral and aural forms (e.g., students discuss in small and whole groups). At the conclusion of each chapter is an expository text summarizing key science concepts, with translations into Spanish and Haitian Creole available on the project website. Additional language development activities and a complete Spanish translation of the curriculum are available on the project website.

Teacher Components of the P-SELL Intervention

Teacher components to support teachers' implementation of the P-SELL curriculum include (a) the teachers' guide and (b) professional development workshops. First, the teachers' guide is designed to assist teachers with curriculum implementation based on the notion of educative curriculum materials (Davis et al., 2014; Davis & Krajcik, 2005; Drake et al., 2014). The front matter of the teachers' guide explains how the curriculum is designed to promote students' mastery of the state science standards, why science inquiry is crucial for facilitating students' understanding of the big ideas in the state science standards, how teachers can help students progress toward student-initiated inquiry, and how teachers can engage all students, especially ELLs, in language and literacy development. For each chapter, following the science inquiry activities, the teachers' guide provides science background information and explanations for the questions under investigation and related natural phenomena, drawing particular attention to students' learning difficulties. In addition, the teachers' guide offers content-specific teaching strategies for each chapter. For example, it provides suggestions on setting up and implementing hands-on activities, along with insights about potential problems and how to respond to such situations. There are suggestions for different levels of guidance and scaffolding by using additional activities for students who need support for content mastery as well as enrichment activities for students who need challenge beyond content mastery. It also includes language

development strategies to promote science learning for students, especially ELLs, and an extensive list of Web-based resources.

Second, the teacher workshops incorporated critical features of effective professional development—content focus, active learning, coherence, sufficient duration, and collective participation (Desimone, 2009; Garet et al., 2001; Penuel, Fishman, Yamaguchi, & Gallagher, 2007). The workshops focused on teachers' science knowledge and reform-oriented teaching practices to promote students' scientific inquiry and understanding while also supporting English language development (i.e., content focus). Teachers were actively engaged in hands-on, science inquiry and planning for classroom implementation, while language development strategies were embedded in promoting students' scientific inquiry and understanding (i.e., active learning). Teachers became familiar with the benchmark clarifications and item specifications of the state science standards that were being assessed at fifth grade (i.e., coherence). The workshops were offered during the summer and throughout the school year, continuing over the three-year period (i.e., duration). Finally, all fifth-grade science teachers within each school and each school district were given time for collaborative planning to develop common goals, share materials, and exchange ideas and experiences arising from the common context of the intervention (i.e., collective participation).

During the first year of the intervention, teachers in the treatment group participated in a summer workshop (or a make-up workshop for those who did not participate in the summer workshop) and workshops during the school year, for a total of five full-day workshops. Workshops were planned collectively among the research team and a district coordinator in each of the three districts. Then, the district coordinators facilitated the workshops in collaboration with the research team. Within this broad work plan, specifics differed across the three school districts in response to each district's policies and guidelines. Teachers received stipends for attending the summer workshop and professional development credits for recertification. When teachers attended workshops during the school year, schools received payments for substitute teachers.

During the first year, the focus of the teacher workshops was on familiarizing the teachers with state science standards and assessment, hands-on inquiry activities and science content, and language development strategies. While engaging in science inquiry activities, teachers discussed science concepts, connected science concepts to one another, and applied science concepts to explain natural phenomena or real-world situations. Additionally, strategies for English language development embedded in science inquiry were introduced. Teachers actively applied the knowledge and strategies they acquired in the workshops throughout the year; they implemented the intervention components in their teaching, reflected on their instructional practices, and shared their experiences and insights with other teachers from the same school and across the schools within the same district.

Method

Research Setting

The study was implemented in three school districts in a Southeastern state. According to the National Center for Education Statistics (2006), District A, located in the northeastern part of the state, is designated as urban. District B, located in the southwestern part of the state, is designated as urban/suburban. District C, located in the central part of the state, is designated as urban/suburban. The three districts encompass a wide range of racial, ethnic, socioeconomic, and linguistic diversity. During the 2012–2013 school year (the first year of our intervention), District A was 45% Black, 8% Hispanic, and 40% White non-Hispanic; 52% participated in a free or reduced price lunch (FRL); and 3% were designated as limited English proficient (LEP, the federal term). District B was 28% Black, 15% Hispanic, 51% White non-Hispanic, 52% FRL, and 8% LEP. District C was 30% Black, 34% Hispanic, 28% White non-Hispanic, 60% FRL, and 14% LEP.

In the state in which the study takes place, all public schools are assigned a "school grade" (A, B, C, D, or F) based on students' performance on high-stakes state assessments. At the elementary level, students are assessed on reading and mathematics in Grades 3 through 5, writing in Grade 4, and science in Grade 5. The percentage of students scoring proficient on the fifth-grade science assessment counts for one-eighth of the school grade.

Research Design

During the 2012–2013 school year, District A had 103 elementary schools, District B had 44 elementary schools, and District C had 125 elementary schools. A cluster randomized controlled trial was conducted. Within each of the three school districts, 22 elementary schools were randomly selected to participate in the study, yielding a total of 66 schools. Because the intervention focused on supporting ELLs in particular, we sampled schools to ensure that the ELL sample in our study would be representative of the participating districts. In each district, 12 schools were randomly selected from a pool of schools with percentages of ELLs higher than the district median, and 10 were selected from a pool of schools with percentages of ELLs lower than the district median. Within each district, 11 of the selected schools were randomly assigned to the treatment group and 11 to the control group, yielding a total of 33 treatment schools and 33 control schools across the three districts. The average school characteristics for the treatment and the control schools are presented in Table 1. There was no significant difference on any of the school characteristics between the two groups at the beginning of the school year.

All fifth-grade science teachers and their students in the 66 schools participated in the study. The teachers and students in the 33 treatment schools

Table 1
Comparison of Treatment and Control School Characteristics

	Treatment (n = 33)		Control (n = 33)		Overall (n = 66)				
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	Difference	*t*	*p*
Percentage of students receiving free/reduced price lunch	71.2	21.7	71.5	20.3	71.3	20.9	0.4	0.07	.94
Percentage of White students	34.1	21.5	33.7	22.2	33.9	21.7	−0.4	−0.07	.94
Percentage of Black students	31.6	28.6	34.9	26.9	33.2	27.6	3.4	0.50	.62
Percentage of Asian students	4.3	5.0	2.9	2.5	3.6	4.0	−1.5	−1.50	.14
Percentage of Hispanic students	26.9	21.6	25.4	18.2	26.1	19.8	−1.5	−0.30	.77
Percentage of exceptional student education students	11.9	3.4	12.3	3.8	12.1	3.6	0.4	0.44	.66
Percentage of English learner students	11.3	12.9	8.5	11.1	9.9	12.1	−2.8	−0.93	.36

Table 2
Time for Science Instruction

Variable	Treatment (%)	Control (%)
Science instruction per week		
Less than 60 minutes	6	4
60–150 minutes	17	25
151–300 minutes	58	58
More than 300 minutes	19	13
Length of science class		
30–45 minutes	26	38
46–60 minutes	54	42
61–75 minutes	17	16
76–90 minutes	2	4

implemented the P-SELL curriculum, whereas those in the 33 control schools implemented the district-adopted textbooks ("business as usual"). District A used *Interactive Science* by Pearson, District B used *National Geographic Science*, and District C used *Science Fusion* by Houghton Mifflin Hartcourt.

Based on questionnaires that teachers filled out at the end of the school year, teachers in both the treatment and control groups taught science regularly and extensively (as shown in Table 2), which reflects the fact that the high-stakes science assessment in this state counts toward school accountability in fifth grade: 58% of the teachers spent 151 to 300 minutes teaching science each week, and another 16% taught science more than 300 minutes each week. Both the total minutes of science instruction per week and the average length of science class were comparable between the treatment and control groups.

Participants

During the 2012–2013 school year, the project involved 123 teachers in the treatment group and 135 teachers in the control group for a total of 258 teachers in the study. There were 6,673 students in the 66 schools at the beginning of the 2012–2013 school year. By the end of the school year, there was no attrition of schools. However, 402 students were no longer in sample schools, 66 students were missing scores on the state science assessment, and 477 students were missing scores on the researcher-developed science assessment at post-administration. Thus, the overall attrition in the first year of implementation was 7.0% for the state science assessment and 13.2% for the researcher-developed science assessment. Differential attrition between the treatment and control groups was 0.8% for the state science assessment and 5.0% for the researcher-developed science assessment.

Multiple imputation using the Markov chain Monte Carlo method (Little & Rubin, 1987) was used to account for missing data at pretest (i.e., multiply

Table 3
Student Demographics by Group

Variable	Group	Treatment (%) (n = 2,894)	Control (%) (n = 3,345)
Gender	Male	47.3	49.0
	Female	52.7	51.0
Ethnicity	Hispanic	28.7	27.7
	Black	24.7	28.4
	White non-Hispanic	37.5	37.1
	Asian	5.3	3.0
	Native American	0.4	0.4
	Mixed	3.4	3.3
Free or reduced price lunch	Received free or reduced price lunch	68.2	67.8
Exceptional student education	Exceptional students	11.2	10.7
English for Speakers of Other Languages	English language learners (ELLs)	9.1	6.6
	Recently reclassified ELLs (exited within 2 years)	3.7	3.6
	Former ELLs (exited over 2 years)	12.5	10.9
	Non-ELLs	74.7	78.8

imputed pretest). Specifically, 20 complete data sets were created based on an imputation model that included all pretest and posttest scores as well as all available demographic variables including gender, race and ethnicity, free and reduced price lunch, language classification, and exceptional status. Twenty data sets were created because a larger number of data sets is considered to be more appropriate when students are nested in classrooms and schools (Lesaux, Kieffer, Kelley, & Harris, 2014). All descriptive and multilevel analyses reported in the following were conducted with the 20 complete data sets and combined using appropriate procedures to aggregate standard errors. As a robustness check, we also conducted the analyses using a data set in which we multiply imputed both the pretest and the posttest and a data set with complete cases only (see Appendix in the online journal). The results were the same using all three approaches: multiply imputed pretest (presented in the following), multiply imputed pretest and posttest, and complete cases only.

The demographics of the fifth-grade students in the treatment and control groups are presented in Table 3. When we included these demographic variables in subsequent models to evaluate the impact of the intervention, the results for the treatment effects were the same as reported in the following.

In the state where the study took place, students are classified into four language proficiency groups: (a) *ELLs* receive services through English for Speakers of Other Languages (ESOL) programs, (b) *recently reclassified ELLs* had exited ESOL services within two years and are monitored for the two-year period, (c) *former ELLs* had exited ESOL services over two years ago, and (d) *non-ELLs* never received ESOL services. In this study, 7.8% of students were ELLs, 3.7% were recently reclassified ELLs, 11.6% were former ELLs, and 76.9% were non-ELLs.

Measures

Student science achievement was assessed using two measures: (a) two equated forms of a researcher-developed science assessment and (b) the high-stakes state science assessment. These two assessments were used for the following reasons. First, the state science assessment is administered once a year at the end of fifth grade, whereas the researcher-developed assessment was administered at the beginning (pre) and the end (post) of the year. The administration of the researcher-developed assessment at the beginning of the year served as a measure of initial science achievement that was used as a covariate in the statistical analyses of the intervention effect (see the data analysis section in the following). Second, the two assessments varied with regard to the degree of alignment to the intervention, thus allowing us to examine whether the intervention produced results that were robust enough to have an effect on the state science assessment as well as the researcher-developed assessment (i.e., proximal vs. distal assessment in Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002; National Research Council [NRC], 2014). For example, the researcher-developed assessment included open-ended items in addition to multiple-choice items to reflect the intervention's focus on language.

Researcher-Developed Science Assessment

Two equated forms of a science assessment were developed to ensure that different forms were used at pretest and posttest each year in an effort to curb the effect of teachers exposing students to items during instruction throughout the school year or students remembering items from the pretest. The two forms were composed of public release items from the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and selected state science assessments. Each form included a spread of low, medium, and high difficulty items across the four bodies of knowledge in the state standards for fifth grade: earth and space science, life science, physical science, and nature of science. Each form had 25 multiple-choice and three short response items. Of the 25 multiple-choice items, 8 remained the same across the two forms

for linking purposes. The three short response items also remained the same on both forms.

For the scoring of the three short response items, we adopted the NAEP scoring rubric that corresponded to each item. A team of raters participated in a two-hour training session prior to scoring responses to each item. All the responses were independently scored by two raters. Disagreements were resolved by a third round of scoring and group consensus, if needed. The interrater agreement for the three items in the pre- and posttest was excellent (weighted kappa above 0.75). The internal consistency, Cronbach's α, of the scores for the pre (form 1) and posttest (form 2) were 0.79 and 0.81, respectively.

The assessment development process included content review, initial tryouts, pilot testing, and item analyses in a manner that is consistent with accepted standards of assessment design (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The initial version of the assessment was piloted the year before implementation with 283 fourth- and fifth-grade students attending six elementary schools in a school district in the same state that was not one of the three districts in the current study.

The researcher-developed assessment was administered to all students in the treatment and control groups at the beginning and the end of the school year by the classroom teachers. Teachers were provided an administration manual with detailed instructions. They were asked to make accommodations for ELLs and students with disabilities according to the guidelines for state assessments.

State Science Assessment

The state science assessment measured student achievement of the state science standards across the four bodies of knowledge: earth and space science, life science, physical science, and nature of science. The assessment included 60 to 66 multiple-choice items administered over the course of two days in 80-minute sessions in April.

Data Analysis

To evaluate whether the intervention had an impact on science achievement of all students (Research Question 1), we fit a sequence of multilevel models (also known as hierarchical linear models) (Raudenbush & Bryk, 2002) in which the score for each measure at the end of the year was regressed on a dummy variable representing condition (treatment or control) and pretest covariates. To improve the precision of the estimate of the treatment effect, we included the pretest score on the researcher-developed assessment at the student level and at the school level as predictors.

The data had a three-level structure whereby students (Level 1) were nested in teachers (Level 2), which in turn were nested in schools (Level 3). Therefore, three-level models were used with two dummy-coded variables for district included at Level 3. To describe these models, we denote the score on the student outcome variable for the *i*th student with the *j*th teacher of the *k*th school by SCORE_{ijk}. The general form of the multilevel model is given as:

$$\text{Level}-1\,\text{model}: \text{SCORE}_{ijk}=\pi_{0jk}+\pi_1\left(\text{PRE}_{ijk}\right)+\text{e}_{ijk}$$

$$\text{Level}-2\,\text{model}: \pi_{0jk}=\beta_{00k}+r_{0jk}$$

$$\begin{aligned}\text{Level}-3\,\text{model}: \beta_{00k}&=\gamma_{000}+\gamma_{001}(\text{TRT}_k)+\gamma_{002}(\text{Dist1}_k)\\&+\gamma_{003}(\text{Dist2}_k)+\gamma_{004}(\text{Pre}_k)+u_{00k},\end{aligned}$$

where PRE_{ijk} corresponds to the student pretest score centered at the school's mean, TRT_k corresponds to school condition (treatment vs. control), Dist1_k and Dist2_k correspond to two dummy-coded variables representing district, Pre_k corresponds to the pretest school mean centered at the district mean, and errors are denoted by e_{ijk}, r_{0jk}, and u_{00k}. Separate multilevel models were estimated for the two outcome variables (i.e., SCORE_{ijk}): (a) posttest scores on the researcher-developed science assessment and (b) end-of-year scores on the state science assessment. The overall treatment effect is represented by the coefficient γ_{001}.

To examine whether the intervention was beneficial for students in each of the language proficiency groups—ELLs, recently reclassified ELLs, former ELLs, and non-ELLs—subgroup analyses were conducted comparing ELLs in the treatment group against ELLs in the control group, recently reclassified ELLs in the treatment group against recently reclassified ELLs in the control group, former ELLs in the treatment group against former ELLs in the control group, and non-ELLs in the treatment group against non-ELLs in the control group. The same multilevel models as previously described were used.

Because the two outcome measures—the researcher-developed assessment and the state science assessment—are related, we used the Benjamini-Hochberg correction to limit cumulative Type I error for the tests of statistical significance on γ_{001} for each outcome (Benjamini & Hochberg, 1995).

Results

Table 4 shows descriptive statistics at pretest and posttest for the researcher-developed assessment and for the state science assessment at the end of the year. The table describes performance overall and by language proficiency groups. Students classified as ELLs had the lowest mean

scores, followed by recently reclassified ELLs. Former ELLs, however, had slightly higher means than non-ELLs, indicating that students who start out as ELLs, after two years or longer since reclassification, can perform comparably to non-ELLs.

Research Question 1: Main Effect of the Intervention

As shown in Table 5, we found significant and meaningfully sized average intervention effects on the researcher-developed science assessment scores (d = 0.25, $p < .001$) and the state science assessment scale scores (d = 0.15, p = .003). This finding indicates that students in the treatment group outperformed students in the control group on both measures of science achievement.

Research Question 2: Subgroup Analyses by Language Proficiency Groups

Subgroup analyses by language classification revealed that the P-SELL intervention had significant and meaningfully sized effects for ELLs (d = 0.35, $p < .001$), recently reclassified ELLs (d = 0.41, p = .020), former ELLs (d = 0.28, $p < .001$), and non-ELLs (d = 0.24, $p < .001$) on the researcher-developed assessment (see Table 6). In other words, ELLs in the treatment group outperformed ELLs in the control group on the researcher-developed assessment, recently reclassified ELLs in the treatment group outperformed recently reclassified ELLs in the control group, former ELLs in the treatment group outperformed former ELLs in the control group, and non-ELLs in the treatment group outperformed non-ELLs in the control group (see Table 6).

Significant intervention effects were found on the state science assessment for non-ELLs (d = 0.16, p = .001) and former ELLs (d = 0.18, p = .015). However, the intervention effects were positive but not statistically significant for recently reclassified ELLs (d = 0.13, p = .58) or ELLs (d = 0.12, p = .247) (see Table 7).

Additional analyses were conducted to investigate whether students' language proficiency moderated the intervention effect by testing interactions between language proficiency group and treatment, using data from the entire sample. Such moderation analysis is a different but complementary approach from subgroup analyses in that it invokes different null hypotheses (i.e., that treatment effects do not differ from one another rather than that each treatment effect does not differ from 0) as well as different assumptions (e.g., that relations between the covariates and outcomes are the same across subgroups). No significant interaction effects were found for either outcome, suggesting little evidence that the treatment impacts differed significantly by language proficiency group (see Table 8).

For the researcher-developed assessment, this result is consistent with results from the subgroup analyses while also suggesting that the apparent differences in the magnitude of the intervention effects among the four

Table 4
Descriptive Statistics on Outcome Measures

	Sample Size		Researcher-Developed Assessment				State Science Assessment	
			Pretest		Posttest			
	Treatment, *n*	Control, *n*	Treatment *M* (*SD*)	Control *M* (*SD*)	Treatment *M* (*SD*)	Control *M* (*SD*)	Treatment *M* (*SD*)	Control *M* (*SD*)
All students	2894	3345	15.52 (5.45)	15.56 (5.48)	19.87 (5.58)	18.47 (5.72)	205.26 (22.47)	202.29 (21.8)
Non-ELLs	2162	2636	16.09 (5.42)	15.93 (5.49)	20.25 (5.5)	18.71 (5.74)	207.48 (21.74)	203.44 (21.43)
Former ELLs	361	365	15.78 (4.96)	16.13 (4.89)	20.44 (4.81)	19.50 (5.09)	207.63 (20.36)	205.68 (20.62)
Recently reclassified ELLs	107	122	13.78 (4.77)	14.01 (4.68)	20.11 (5.77)	17.87 (5.45)	203.15 (20.77)	200.48 (21.28)
ELLs	264	222	10.86 (4.22)	10.73 (4.08)	15.76 (5.56)	14.29 (4.91)	184.56 (21.23)	184.32 (20.13)

Note. ELL = English language learner.

Table 5
Main Effect of the Intervention (Analyses With Imputed Pretest Data)

	Researcher-Developed Assessment	State Science Assessment
n		
School	66	66
Teacher	252	258
Student	5,752	6,173
Fixed effects		
Intercept	18.22***	200.34***
Treatment	1.41***	3.28**
Pretest (school average)	1.09***	4.13***
Pretest (student level)	0.69***	2.69***
District 1	−0.35	−0.38
District 2	0.75**	4.20***
Random effects		
School	0.13	4.76*
Teacher	1.31***	35.21***
Student	13.52***	179.50***
Effect size		
Treatment	0.25	0.15

Note. The maximum score on the researcher-developed test is 31 points. The state science assessment score scale ranges from 140 to 260.
*$p < .05$. **$p < .01$. ***$p < .001$.

subgroups may not be meaningful. For the state science assessment, this result is somewhat contrary to the results from the subgroup analyses in that this result suggests that the ELLs and recently reclassified ELLs may have benefitted similarly from the intervention relative to the other students. However, given differences in the null hypotheses and assumptions in the two approaches, we take the conservative position of emphasizing the subgroup analyses as indicating that the intervention was effective in improving state science assessment scores for non-ELLs and former ELLs but may not have been effective in improving these scores for ELLs and recently reclassified ELLs.

Discussion and Implications

This study evaluated the effects of P-SELL, a curricular and professional development intervention designed to improve science achievement of fifth-grade students with a focus on ELLs. The P-SELL intervention was built on previous research demonstrating the promise of integrated models of instruction to promote inquiry-based science while supporting English language development (Buxton & Lee, 2014; Janzen, 2008; Lee, 2005). We examined the impact of the P-SELL intervention during the first year of a large-scale implementation across three school districts in one state.

Table 6

Main Effect of the Intervention on the Researcher-Developed Assessment by Language Proficiency Subgroups (Analyses With Imputed Pretest Data)

Term	All Students	Non-ELLs	Former ELLs	Recently Reclassified ELLs	ELLs
n					
School	66	66	57	46	55
Teacher	252	246	197	112	126
Student	5,752	4,406	688	207	451
Fixed effects					
Intercept	18.21***	18.28***	18.37***	17.97***	16.00***
Treatment	1.42***	1.36***	1.39***	2.29*	1.86***
Pretest (school average)	1.09***	1.09***	0.97***	1.19***	1.03***
Pretest (student level)	0.69***	0.69***	0.62***	0.65***	0.66***
District 1	−0.35	−0.44	0.41	1.98	1.61*
District 2	0.75**	0.63**	1.63***	1.27	1.81**
Random effects					
School	0.12	0.09	0.02	0.18	0.10
Teacher	1.33***	1.51***	0.92	3.53	1.30
Student	13.57***	13.26***	10.54***	15.77***	16.74***
Treatment effect size	0.25	0.24	0.28	0.41	0.35

Note. ELL = English language learner.
*$p < .05$ **$p < .01$ ***$p < .001$

Discussion

The multilevel analyses revealed that the P-SELL intervention had a positive impact on students' science achievement as measured by both the researcher-developed science assessment and the state science assessment in the first year of implementation. Maerten-Rivera et al. (in press) also found a positive impact on the state science assessment; however, an effect was not found until the second year of implementation. In the current study, the impact was both statistically significant and practically important as evidenced by the magnitude of the effects: 0.25 on the researcher-developed assessment and 0.15 on the state science assessment. According to Lipsey et al. (2012), the mean effect size of interventions that focus on curriculum or broad instructional programs is 0.13, and the median effect size is 0.08. Thus, effect sizes of 0.15 on the state science assessment and 0.25 on the researcher-developed assessment are of practical importance. Another way to interpret the effect sizes is in relation to the average year-to-year growth in science based on national norms (Hill, Bloom, Black, & Lipsey, 2008; Lipsey et al., 2012). Students gain about 0.40 standard deviations on nationally normed standardized science tests between the spring of fourth and fifth

Table 7

Main Effect of the Intervention on the State Science Assessment by Language Proficiency Subgroups (Analyses With Imputed Pretest Data)

Term	All Students	Non-ELLs	Former ELLs	Recently Reclassified ELLs	ELLs
n					
School	66	66	58	46	55
Teacher	258	252	202	116	134
Student	6,173	4,743	721	228	481
Fixed effects					
Intercept	200.31***	200.99***	200.32***	197.88***	190.22***
Treatment	3.33**	3.41**	3.64*	2.73	2.49
Pretest (school average)	4.13***	4.04***	3.88***	3.52***	3.88***
Pretest (student level)	2.69***	2.66***	2.56***	2.61***	2.65***
District 1	−0.40	−0.82	3.58	11.05***	7.06
District 2	4.22***	3.50**	8.26***	8.70***	9.62***
Random effects					
School	4.22	3.97	4.24	1.18	5.49
Teacher	35.59***	33.89***	28.98	13.24	37.03
Student	180.15***	174.15***	151.66***	231.90***	215.26***
Treatment effect size	0.15	0.16	0.18	0.13	0.12

Note. ELL = English language learner.
*$p < .05$ **$p < .01$ ***$p < .001$

grade (Lipsey et al., 2012). Assuming that the year-to-year growth in the districts in our study is comparable to the national norms, an effect size of 0.15 on the state science assessment represents about a 38% improvement over the annual gain otherwise expected.[1]

The fact that an effect of practical importance was found not only on a researcher-developed assessment but also on the state science assessment is noteworthy considering that the P-SELL intervention was implemented large-scale across three school districts and under routine conditions. It is also noteworthy considering that science was being taught extensively in both treatment and control schools. In a review of impact studies of elementary science programs, Slavin et al. (2014) found that studies of inquiry-oriented programs that provided science kits did not result in positive achievement impacts and that the weighted overall mean effect size across the six studies of science kit programs was only 0.02. They also found that inquiry-oriented professional development programs that did not provide kits showed positive science achievement outcomes, with a weighted mean effect size of 0.36. However, none of these professional development studies were of the scale of the P-SELL intervention (in terms of numbers of schools and students), and none of the interventions focused on ELLs. Our

Table 8
Language Proficiency Group as a Moderator of the Intervention Effect (Analyses With Imputed Pretest Data)

	Researcher-Developed Posttest	State Science Assessment
N		
School	66	66
Teacher	252	258
Student	5,752	6,173
Fixed effects		
Intercept	18.20***	200.36***
Treatment	1.36***	3.40**
Pretest (school average)	1.11***	4.10***
Pretest (student level)	0.68***	2.62***
District 1	−0.27	−0.08
District 2	0.84***	4.89***
Former English language learners (ELLs; school percentage)	0.03	6.05
Recently reclassified ELLs (school percentage)	−1.99	−12.62
ELLs (school percentage)	1.91	5.54
Former ELLs (student level)	0.49*	1.28
Recently reclassified ELLs (student level)	0.26	1.18
ELLs (student level)	−1.27***	−6.88***
Treatment × Former ELLs	−0.08	−0.13
Treatment × Recently Reclassified ELLs	0.73	−0.09
Treatment × ELLs	0.30	−0.39
Random effects		
School	0.12	4.18*
Teacher	1.36***	34.45***
Student	13.44***	177.37***
Effect sizes		
Treatment × Former ELLs	−0.02	−0.01
Treatment × Recently Reclassified ELLs	0.14	−0.01
Treatment × ELLs	0.07	−0.02

intervention provides both science supplies (kits) for inquiry activities and professional development for teachers. The findings of this study provide strong experimental evidence that an intervention that promotes science inquiry and language development for ELLs can be scaled up, implemented across varied educational settings, and result in improvements for *all* students not only on a researcher-developed assessment but on the mandated state assessment used for school accountability.

We also examined whether the P-SELL intervention had positive effects on students of varying levels of English proficiency, including ELLs, recently reclassified ELLs, former ELLs, and non-ELLs. Using subgroup analyses, we found that the intervention had a positive and significant effect for each of the subgroups on the researcher-developed assessment. The P-SELL intervention also had a positive and significant effect for non-ELLs and former ELLs on the state science assessment. However, the intervention had positive but not statistically significant effects for ELLs and recently reclassified ELLs on the state science assessment.

One explanation for these findings may have to do with the proximal versus distal relationship of the assessments to the intervention (NRC, 2014; Ruiz-Primo et al., 2002). The researcher-developed assessment was not a treatment-inherent measure in that it was designed to assess the fifth-grade science standards that all fifth-grade students, not just those in the treatment schools, would have been exposed to. Also, the assessment was composed of NAEP and TIMSS items, not items specifically developed to align with the intervention. However, in developing the assessment, care was taken to ensure that the science content assessed by the items selected was covered by the intervention. Also, items with less frequently occurring vocabulary words that were unrelated to the science content being assessed and could be unfamiliar to ELLs (e.g., *cupboard*) were not selected.

Another explanation may have to do with the different nature of the assessments. The researcher-developed assessment included three short-response items that may have given ELLs and recently reclassified ELLs the opportunity to better show their science understanding than the state science assessment that consisted exclusively of multiple-choice items. Goldschmidt, Martinez, Niemi, and Baker (2007) found evidence that open-ended responses might be less affected by student background variables, including ELL status, than those on multiple-choice tests. Abedi (2010) further argues that "open-ended assessments improve the chances for ELL students to engage with language production and learning, offering unique opportunities for ELL students to express their knowledge in a broader sense than the limited linguistic opportunities given to them in traditional multiple choice items" (p. 4). This raises questions about the exclusive use of multiple-choice tests for assessing ELLs and recently reclassified ELLs.

Future Research

In this article, we report the results of the first year of a three-year implementation. Using data from all three years of the P-SELL intervention's implementation, we will be able to examine the impact of the intervention over time in the same schools with the same teachers across three different cohorts of students. Professional development researchers have noted delayed effects of professional development activities (Kreider & Bouffard,

2006). For example, Silverstein, Dubner, Miller, Glied, and Loike (2009) reported that differences in passing rates on high-stakes science assessments following teacher professional development between students of participating teachers and non-participating teachers were not significant during the first two years but were significant during the third and fourth years. As mentioned earlier, Maerten-Rivera et al. (in press) found an impact of the intervention in the second and third years but not the first year. Using data from three years of implementation, we will be able to examine whether the positive impacts of the intervention in its first year can be sustained over time and whether there are any differential effects for students of varying levels of English proficiency over time. In addition, an important next step—should positive outcomes continue to be observed—would be a careful consideration of program costs, as successful implementation will require school districts to invest in key features of the intervention, including professional development, curriculum materials, and science supplies.

Contributions

Zwiep and Straits (2013) argue that "what is needed to address this (pandemic) failure to develop the science literacy of countless ELLs is a body of research that demonstrates the merits of high-quality, inquiry science instruction for science achievement and second language acquisition" (p. 1317). This study makes valuable contributions, both methodological and conceptual, toward building this body of research. The study also makes a valuable contribution to the literature on impact studies of elementary science programs in general.

Methodologically, this study had several strengths. First, the study employed a randomized controlled trial design that has strong internal validity and allows for a causal interpretation of the impact of the intervention. Second, by administering the researcher-developed science assessment at the beginning and the end of the school year, the study was able to include a covariate to improve the precision of the treatment effects. This additional measure is particularly important in science because unlike high-stakes assessments in reading and mathematics administered at Grades 3, 4, and 5, science is typically administered once in elementary school (typically fourth or fifth grade) and thus performance from previous grades is not available as a control variable. Third, the study involved randomly selected schools across three demographically diverse school districts in one state. Random selection of schools in varied educational settings enhances generalizability of the findings and allows the examination of whether the intervention effect replicates across a set of diverse school districts. Fourth, the intervention involved all fifth-grade science teachers in the participating schools rather than a self-selected group of volunteer teachers, also enhancing the generalizability of the findings. Collective participation of teachers in

professional development opportunities is critically important for elementary school teachers who often lack adequate preparation in science or those teachers who do not consider teaching for diversity as their responsibility (Banilower et al., 2013). Fifth, the study included groups of varying levels of English proficiency, not just ELLs and non-ELLs as most studies do. By including recently reclassified and former ELLs, the study provides a more nuanced and complete picture of the role of language proficiency in the effectiveness of an intervention. Finally, as a scale-up of an intervention model, the study was implemented across varied educational settings and under typical conditions in elementary science classrooms. This scale-up context requires implementation of a science program within the participating districts' constraints of science instructional time, professional development in science with student diversity including ELLs, a science curriculum that meets the learning needs of diverse student groups including ELLs, and science supplies and equipment essential for quality science instruction. Thus, the results of the study contribute to the emerging literature on scale-up research (McDonald, Keesler, Kauffman, & Schneider, 2006; Schneider & McDonald, 2007a, 2007b), especially in science education (Lee & Luykx, 2005; see also Lee & Krajcik, 2012).

Conceptually, our intervention provides equitable learning opportunities for ELLs while conceptualizing essential elements of high-quality science education for *all* students. The key features of the P-SELL intervention (standards-based, inquiry-oriented, and language-focused) were consistently delivered via educative curriculum materials (Davis et al., 2014; Davis & Krajcik, 2005; Drake et al., 2014) and key features of effective professional development (Desimone, 2009; Garet et al., 2001; Penuel et al., 2007). In addition, the P-SELL intervention enabled students of varying levels of English proficiency to engage in science inquiry and develop scientific understanding while meeting the demands of high-stakes science assessment and accountability policy. By showing a positive impact on both the researcher-developed and the state high-stakes science assessment, this study demonstrated via an experimental design that inquiry-based approaches can improve student achievement even in the context of high-stakes assessment and accountability policy. The study further demonstrated that inquiry-based approaches can be beneficial for all students, including those of varying levels of English proficiency.

As ELLs are the fastest growing student population in the nation, their academic success in both content and language is critical for their participation in college, careers, and citizenship in U.S. society and the global community. The results of this study, supported by the emerging literature on science education with ELLs, indicate that one way to facilitate their success is through educational interventions that promote inquiry-based science. As states begin to implement curricula aligned to the NGSS or the principles of effective science instruction in the *Framework for K–12 Science Education*

(National Research Council, 2012) from which the NGSS were developed, it will become important to investigate whether interventions focusing on three-dimensional science learning and the language-intensive science and engineering practices in particular (Lee et al., 2013) are similarly beneficial to the ELL population.

Notes

This work is supported by the National Science Foundation (NSF Grant DRL 1209309). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the position, policy, or endorsement of the funding agency.

[1]The state where the study was conducted does not test science in fourth grade, and thus data on year-to-year growth in science specific for that state are not available.

References

Abedi, J. (2010). *Performance assessments for English language learners*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: APA.

August, D., Branum-Martin, L., Cárdenas-Hagan, E., Francis, D., Powell, J., Moore, S., & Haynes, E. (2014). Helping ELLs meet the Common Core State Standards for literacy in science: The impact of an instructional intervention focused on academic language. *Journal of Research on Educational Effectiveness*, *7*(1), 54–82.

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

Banilower, E. R., Smith, P. S., Weiss, I. R., Malzahn, K. A., Campbell, K. M., & Weis, A. M. (2013). *Report of the 2012 National Survey of Science and Mathematics Education*. Chapel Hill, NC: Horizon Research, Inc.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B Methodological*, *57*(1), 289–300.

Buxton, C. A., & Lee, O. (2014). English language learners in science education. In N. G. Lederman & S. K. Abell (Eds.), *Handbook of research in science education* (2nd ed., pp. 204–222). Mahwah, NJ: Lawrence Erlbaum Associates.

Davis, E., & Krajcik, J. (2005). Designing educative curriculum materials to promote teacher learning. *Educational Researcher*, *34*(3), 3–14.

Davis, E. A., Palincsar, A. S., Arias, A., Bismack, A., Marulis, L., & Iwashyna, S. (2014). Designing educative curriculum materials: A theoretically and empirically driven process. *Harvard Educational Review*, *84*(1), 24–52.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181–199.

Drake, C., Land, T. J., & Tyminski, A. M. (2014). Using educative curriculum materials to support the development of prospective teachers' knowledge. *Educational Researcher*, *43*(3), 154–162.

Echevarria, J., Richards-Tutor, C., Canges, R., & Francis, D. (2011). Using the SIOP model to promote the acquisition of language and science concepts with English learners. *Bilingual Research Journal*, *34*(3), 334–351.

Fathman, A. K., & Crowther, D. T. (Eds.). (2006). *Science for English language learners: K–12 classroom strategies*. Arlington, VA: National Science Teachers Association.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? From a national sample of teachers. *American Educational Research Journal*, *38*(4), 915–945.

Goldschmidt, P., Martinez, J. F., Niemi, D., & Baker, E. L. (2007). Relationship among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment*, *12*(3&4), 239–266.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Janzen, J. (2008). Teaching English language learners in the content areas. *Review of Educational Research*, *78*(4), 1010–1038.

Kreider, H., & Bouffard, S. (2006). Questions and answers: A conversation with Thomas R. Guskey. *The Evaluation Exchange*, *XI*(4). Retrieved from http://www.hfrp.org/evaluation/the-evaluation-exchange/issue-archive/professional-development/a-conversation-with-thomas-r.-guskey

Lara-Alecio, R., Tong, F., Irby, B. J., Guerrero, C., Huerta, M., & Fan, Y. (2012). The effect of an instructional intervention on middle school English learners' science and English reading achievement. *Journal of Research in Science Teaching*, *49*(8), 987–1011.

Lee, O. (2005). Science education and English language learners: Synthesis and research agenda. *Review of Educational Research*, *75*(4), 491–530.

Lee, O., Deaktor, R., Enders, C., & Lambert, J. (2008). Impact of a multiyear professional development intervention on science achievement of culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching*, *45*(6), 726–747.

Lee, O., & Fradd, S.H. (1998). Science for all, including students from non-English language backgrounds. *Educational Researcher*, *27*(4), 12–21.

Lee, O., & Krajcik, J. (Eds.). (2012). Large-scale interventions in science education for diverse student groups in varied educational settings [Special issue]. *Journal of Research in Science Teaching*, *49*(3).

Lee, O., & Llosa, L. (2011–2015). *Promoting science among English language learners (P-SELL) scale-up (Discovery Research K-12; NSF Grant DRL 1209309)*. New York, NY: New York University.

Lee, O., & Luykx, A. (2005). Dilemmas in scaling up innovations in science instruction with nonmainstream elementary students. *American Educational Research Journal*, *42*(3), 411–438.

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English language arts and mathematics. *Educational Researcher*, *42*(4), 223–233.

Lesaux, N. K., Kieffer, M. J., Kelley, J. G., & Harris, J. R. (2014). Effects of vocabulary instruction for linguistically diverse adolescents: Evidence from a randomized field trial. *American Educational Research Journal*, *51*(6), 1159–1194.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-

3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: John Wiley.

Maerten-Rivera, J., Ahn, S., Lanier, K., Diaz, J., & Lee, O. (in press). Effect of a multiyear intervention on science achievement of all students including English language learners. *The Elementary School Journal*.

McDonald, S-K., Keesler, V. A., Kauffman, N. J., & Schneider, B. (2006). Scaling-up exemplary interventions. *Educational Researcher*, *35*(3), 15–24.

National Center for Education Statistics. (2006). *Data local education agency locale code file: School year 2005–06*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Center for Education Statistics. (2012). *The nation's report card: Science 2011* (NCES 2012-465). Washington, DC: U.S. Department of Education.

National Center for Education Statistics. (2014). *The condition of education 2014* (NCES 2014-083). Washington, DC: U.S. Department of Education.

National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, *44*(4), 921–958.

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

Rosebery, A. S., & Warren, B. (Eds.). (2008). *Teaching science to English language learners: Building on students' strengths*. Arlington, VA: National Science Teachers Association.

Rosebery, A. S., Warren, B., & Conant, F. R. (1992). Appropriating scientific discourse: Findings from language minority classrooms. *Journal of the Learning Sciences*, *2*(1), 61–94.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369–393.

Santau, A. O., Maerten-Rivera, J. L., & Huggins, A. C. (2011). Science achievement of English language learners in urban elementary schools: Fourth-grade student achievement results from a professional development intervention. *Science Education*, *95*(5), 771–793.

Saunders, W. M., & Marcelletti, D. J. (2013). The gap that can't go away: The catch-22 of reclassification in monitoring the progress of English language learners. *Educational Evaluation and Policy Analysis*, *35*(2), 139–156.

Schneider, B., & McDonald, S. K. (Eds.). (2007a). *Scale-up in education: Ideas in principle* (Vol. 1). New York, NY: Rowman & Littlefield Publishers.

Schneider, B., & McDonald, S. K. (2007b). *Scale-up in education: Issues in practice* (Vol. 2). New York, NY: Rowman & Littlefield Publishers, Inc.

Silverstein, S. C., Dubner, J., Miller, J., Glied, S., & Loike, J. D. (2009). Teachers' participation in research programs improves their students' achievement in science. *Science*, *326*(5951), 440–442.

Slavin, R.E., Lake, C., Hanley, P., & Thurston, A. (2014). Experimental evaluations of elementary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching*, *51*(7), 870–901.

U.S. Census Bureau. (2012). *Statistical abstract of the United States, 2012*. Washington, DC: Government Printing Office.

Zwiep, S., & Straits, W. (2013). Inquiry science: The gateway to English language proficiency. *Journal of Science Teacher Education*, *24*(8), 1315–1331.

Zwiep, S., Straits, W. J., Stone, K. R., Beltran, D. D., & Furtado, L. (2011). The integration of English language development and science instruction in elementary classrooms. *Journal of Science Teacher Education*, *22*(8), 769–785.

Manuscript received May 7, 2015
Final revision received October 1, 2015
Accepted October 30, 2015