LANGUAGE
TEACHING
RESEARCH

Language Teaching Research
15(3) 363–380

DOI: 10.1177/1362168811401155
ltr.sagepub.com
SAGE

Research on academic literacy development in sheltered instruction classrooms

Deborah J. Short
Center for Applied Linguistics, Washington, DC, USA

Jana Echevarría and Catherine Richards-Tutor
California State University, USA

Abstract

This article describes an extended program of research in sheltered instruction and the effects on the academic literacy development of English language learners. It also highlights the challenges of scaling up an instructional intervention. The intervention was the Sheltered Instruction Observation Protocol (SIOP) Model, an approach that teaches subject area curriculum to students learning through a second language using techniques that make the content material accessible and also help develop the students' second language skills. Three successive studies looked at teacher change over time and student achievement on standardized assessments and researcher-developed measures. Results of the three studies reveal that students with teachers who were trained in the SIOP Model of sheltered instruction and implemented it with fidelity performed significantly better on assessments of academic language and literacy than students with teachers who were not trained in the model. The article offers guidance for strengthening professional development for teachers so the quality of instruction they deliver to English language learners improves and the students strengthen their English language and academic outcomes.

Keywords

classroom research, second language teacher development, sheltered content instruction, academic literacy

Corresponding author:
Deborah J. Short, Center for Applied Linguistics, 5719 South Third Street, Arlington, VA 22204, USA
Email: dshort@cal.org

I Introduction

'Sheltered instruction' in schools in the USA generally refers to a subject class such as mathematics, science, or history taught through English wherein many or all of the students are second language learners. This article relates the story of an extended program of research in sheltered instruction and the effects on the academic literacy development of English language learners (ELLs). It also highlights the challenges of scaling up an instructional intervention. The intervention was the Sheltered Instruction Observation Protocol (SIOP) Model, an approach that teaches subject area curriculum to students learning through a second language using techniques that make the content material accessible and also helps develop the students' second language skills. Results of the three studies discussed here reveal that students with teachers who are trained in the SIOP Model and implement it with fidelity perform better on assessments of academic language and literacy than students with teachers who are not trained in the model.

The SIOP Model was developed initially for content teachers of students learning the subject matter through their second language. It evolved also as an approach for teachers of English to use to integrate content material (e.g. subject matter vocabulary, expository reading passages) in their lessons. Therefore, it is an approach for integrating language and content instruction in either content areas or language development classes.

II The CREDE SIOP Model development study

The SIOP Model was developed through a seven-year research study (1996–2003), 'The effects of sheltered instruction on the achievement of limited English proficient students', for the Center for Research on Education, Diversity & Excellence (CREDE) and funded by the US Department of Education. Researchers from California State University Long Beach and the Center for Applied Linguistics collaborated with middle-school teachers in three metropolitan districts in the USA (east and west coasts) to build and test a model of sheltered instruction. An observation tool for researchers to measure teachers' implementation of sheltered instruction was developed first, the Sheltered Instruction Observation Protocol (SIOP). During the first four years of the study, the teachers field-tested variations of the model in their classrooms and the researchers monitored the effects. By 2000, the SIOP had grown into a lesson planning and delivery approach known as the SIOP Model. It has 30 features of instruction grouped into eight components: lesson preparation, building background, comprehensible input, strategies, interaction, practice & application, lesson delivery, and review & assessment (Echevarría et al., 2008).

The SIOP Model is a framework for teachers to present curricular content concepts to ELLs through strategies and techniques that make information comprehensible to the students. While doing so, teachers develop student academic language skills across the domains of reading, writing, listening, and speaking. The model combines features recommended for high quality instruction for all students, such as cooperative learning and reading comprehension strategies (Genesee et al., 2006) with specific features for second language learners, such as language objectives, oral language practice, and academic vocabulary development. The model allows for some natural variation in teaching styles and lesson delivery but attention to academic literacy is required. Teachers may accomplish their language and content goals in multiple ways suited to a particular lesson, such as

engaging students in peer discussions about a science experiment using specific language frames or holding a class debate about the impact of immigration on the economy.

1 SIOP instrument reliability and validity study

We formalized the observation protocol with a 5-point scale for each feature on the SIOP Model, so the level of implementation could be measured for any lesson. Researchers and coaches could rate teachers' lessons over time using this protocol and provide explicit feedback to help them implement the model with more fidelity. A separate study with independent raters established the SIOP protocol as a highly reliable and valid measure of sheltered instruction (Guarino et al., 2001).

2 SIOP writing assessment study

After finalizing the SIOP Model as a lesson delivery system and ensuring teachers could teach it with high fidelity, we investigated the model's effects on student academic language achievement in a small quasi-experimental study with 19 treatment teachers and four comparison teachers. The main research question was:

> Are there significant differences in achievement data for students of treatment teachers who receive SIOP training vs. students in sheltered classes whose teachers have not received SIOP training?

At that time, most ELLs in our districts were exempted from the standardized testing process. So, we used the Illinois Measurement of Annual Growth in English (IMAGE) writing assessment as an outcome measure of academic literacy. This Illinois standardized test was developed specifically for ELLs and measured their annual growth in reading and writing skills for Grades 3 (mainly 8-year-olds) and higher. The test was valid and reliable and had correlational and predictive value for English language learners' scores on the state standardized assessments of reading and mathematics designed for native English-speaking students (Illinois State Board of Education, Assessment Division, 2004). It had five subtests: language production, focus, support/elaboration, organization, and mechanics.

During the 1998–99 school year, we administered the IMAGE writing pre-test to middle-school ELLs in the fall and the post-test in the spring. Participants were ELLs in sheltered classes whose teachers were trained in the SIOP Model (treatment group, $n = 241$) and ELLs in the same district programs whose teachers had no exposure to the SIOP Model (comparison group, $n = 77$). Students in both groups were in Grades 6–8 (10–14-years-olds), with mixed English proficiency levels and over 10 different native languages.

Mean score analyses of the pre- and post-test writing samples (for total score and subtest scores) revealed that treatment students performed less well on all pre-tests compared to the comparison students but significantly better on the post-tests (Echevarría et al., 2006); see Table 1.

Because of these differences, analyses of co-variances (ANCOVA) were conducted with adjusted post-test means to determine whether treatment students made statistically significant gains in writing compared to the other students. The ANCOVA results in Table 2 show significant main effects for the treatment condition on the total score and on three of the five

Table 1 CREDE SIOP: Writing mean scores for treatment and comparison groups

	Treatment		Comparison	
	Pre-test	Post-test	Pre-test	Post-test
Language production:				
M	2.65	3.22	2.77	3.09
SD	.78	.79	.78	.73
n	240	240	77	77
Focus:				
M	2.81	3.30	3.01	3.17
SD	.87	.98	.88	.94
n	239	239	77	77
Support/elaboration:				
M	2.65	3.26	2.83	3.18
SD	.78	.72	.70	.81
n	241	241	77	77
Organization:				
M	2.77	3.31	3.16	3.21
SD	.96	.78	.92	.71
n	241	241	77	77
Mechanics:				
M	2.72	3.28	2.84	3.17
SD	.88	.87	.86	.94
n	241	241	77	77
Total:				
M	13.55	16.36	14.61	15.81
SD	3.42	3.33	3.36	3.45
N	238	238	77	77

subtests. Participants whose teachers were trained in the SIOP Model made significantly greater gains than the comparison group in the total writing score ($p = .001$) and on language production ($p = .026$), organization ($p = .018$), and mechanics ($p = .044$). The treatment group also made gains over the comparison group in the focus and support/elaboration subtests, but the gains did not reach statistical significance (Echevarría et al., 2006).

We also calculated the effect size (Cohen's *d*) of the intervention, which was .833. This effect size is considered large by most indices (Cohen, 1998), suggesting significant gains over time in students' overall writing performance as a result of the SIOP intervention.

3 Discussion

The CREDE study revealed positive effects of the SIOP Model on student literacy achievement as measured with the IMAGE writing assessment. The treatment group made greater gains during the school year, increasing an average of 2.9 points between pre-test and post-test administrations compared to an average gain of 0.7 points for the comparison group, based on adjusted post-test means. Although two subtest gains did not reach significance, the overall results were persuasive, given that none of the classes

Table 2 Analysis of covariance of post-test writing results by treatment condition

Variable	*M* square	*F*-ratio	*p*
Language production	2.133	5.004	0.026*
Focus	2.904	3.706	0.055
Support/elaboration	1.247	2.680	1.030
Organization	2.842	5.651	0.018*
Mechanics	2.065	4.101	0.044*
Total	78.276	10.785	0.001*

Note: Pre-test scores served as the covariate for post-test dependent measures. * $p < .05$

were English as a second language (ESL) classes where writing is directly instructed. These results indicated that the SIOP Model offered a promising approach for helping English language learners develop academic literacy skills needed for success in school, in this case academic writing.

4 SIOP professional development

In the final two years of CREDE, with additional support from the Fund for the Improvement of Postsecondary Education, researchers designed a professional development program to help teachers learn and implement the model. The resources included a professional development manual (Short et al., 2002), and two videos demonstrating the model in classrooms with varied grade levels, English proficiency levels, and subject areas (Hudec & Short, 2002a, 2002b). These professional development resources were utilized extensively in later research studies.

III The New Jersey SIOP quasi-experimental study

The next step in the SIOP program of research was a larger quasi-experimental study, 'Academic literacy through sheltered instruction for secondary English language learners', to determine SIOP Model effects on middle-school and high-school student performance. It also examined the SIOP professional development intervention offered in the 2004–05 and 2005–06 school years. Funded by the Carnegie Corporation of New York and the Rockefeller Foundation, it was conducted by researchers at the Center for Applied Linguistics. The study focused on two main research questions:

- Will English language learners in one district with teachers who received professional development in the SIOP Model show significantly higher achievement in reading, writing, and oral proficiency in English on a standardized measure than ELLs in a comparable district with teachers who had no SIOP Model professional development?
- Do teachers reach high levels of implementation of the SIOP Model during a sustained professional development program after one or two years?

The study took place in two matched districts (one treatment, one comparison) in northern New Jersey and compared groups of teachers and students in the secondary schools and analysed student achievement on state standardized tests. Each district has two middle schools and one high school. The outcome measure of academic literacy was the IDEA Language Proficiency Tests (IPT), the state standardized assessment of English language proficiency. We also looked at student performance on several state content achievement tests and measured teacher fidelity to SIOP implementation using the SIOP protocol.

We recruited a representative sample of teachers in both districts who taught in Grades 6–12. More than half of the teachers were veterans, with 10 years or more of experience. Most were voluntary participants, although a few were assigned to the project.

Two cohorts of teachers joined were trained in the treatment district. Cohort 1 began in Year 1 (2004–05) and 35 teachers participated for two years. Cohort 2 joined in Year 2 (2005–06) with 23 teachers. Treatment teachers taught mathematics, science, history, language arts, ESL, special education, and technology. The treatment district also had three coaches in the first year and five in the second. All the coaches were current or former ESL teachers and had been trained in the SIOP Model. The comparison district did not have cohorts. Nineteen teachers participated both years. The comparison teachers taught mathematics, science, history, and ESL.

Students who were in the ESL programs in Grades 6–12 in both districts were included in our data collection. Treatment students (n = 387) spoke more than 15 different native languages and were from 35 countries of origin. Comparison students (n = 193) spoke eight different native languages with 25 countries of origin. Spanish, Polish, Arabic, and Portuguese were the most common languages.

Administrative interest was strong in the treatment district. The superintendent was supportive of the project and stayed informed of the progress as did the school principals. The ESL/bilingual director was involved in many ways, from securing funding for the coaching assignments to hand-scheduling high-school ELLs so that more would have classes with SIOP-trained teachers.

1 SIOP intervention professional development

The professional development program in the treatment district included (1) workshops for coaches and teachers, (2) classroom observations and coaching, and (3) technical assistance via electronic media. The program focused on instructional strategies for teaching academic English literacy skills and content knowledge to ELLs through the SIOP Model. It utilized the print and video resources developed during the CREDE study. Researchers used a participatory approach with modeling, hands-on activities, cooperative mini-projects, analysis of videotaped instruction, and integration of research and theory to help teachers incorporate the model into their teaching. Each cohort had seven days of professional development in its first year of participation. Cohort 1 had three additional days in the second year.

Coaches and researchers observed and gave feedback to teachers to assist with implementation. Because the coaches were on-site, some teachers also sought advice from them informally. The project website posted sample lesson plans and step-by-step explanations of instructional techniques. Teachers could participate in online chats with the

researchers and use the closed group electronic list to share information, challenges, and successes. Some teachers took more advantage of the coaching and technical support than others.

Comparison teachers did not receive SIOP Model professional development. They participated in regular district trainings two or three days each year. Each school also had a one-hour workshop to discuss student diversity and accommodations for ELLs in classrooms. ESL teachers had additional workshops on topics such as designing thematic units, using the new content-based ESL textbooks, and demonstrating instructional techniques.

2 Data collection and analysis of teacher implementation

Knowing that sheltered instruction has been addressed in ESL teacher education for 20 years (Crandall, 1993; Short, 2006), we anticipated that comparison teachers might incorporate some characteristics of SIOP lessons in their classrooms. We therefore observed, took field notes, and rated lessons of each participating treatment and comparison teacher twice per year, in the fall and spring, to determine whether features of the SIOP Model were being implemented. Using the SIOP protocol's rating scale, we scored the lessons and assessed teachers' fidelity to the intervention and how their level of implementation changed over time. Low implementers scored 50% or below on the scale; medium implementers scored between 50% and 75%; and high implementers scored 75% or higher.

Teacher implementation data revealed that treatment teachers incorporated more features of sheltered instruction than comparison teachers. At the treatment site, after one year of professional development, 56% of Cohort 1 and 74% of Cohort 2 were high implementers of the SIOP Model. After two years, 71% of Cohort 1 reached a high level. At the comparison site, only 5% of the teachers reached a high level in the first year; only 17% by the second year. The features of the SIOP Model were thus implemented more extensively in the treatment district (Short et al., to appear).

3 Data collection and analysis of student achievement

a English language proficiency: We collected the IPT oral language, reading, writing and total English proficiency scores for ELLs in both districts. First, we gathered baseline IPT data on all ELLs from the spring 2004 administration, before the study began. In 2005 and 2006, in the treatment district, we collected IPT scores of ELLs with at least one SIOP teacher. In the comparison district, we collected the IPT scores of ELLs with participating teachers.[1] In these district ESL programs, new students enter and others exit annually. As a result we had a cross-section of students that was not matched across the years and so we examined the average mean scores of the groups. Because the districts had a high level of student mobility, only a small number participated in all three IPT administrations, so no longitudinal analyses were undertaken.

We compared IPT mean proficiency level scores for treatment and comparison groups each year. Then, using Year 2 data, we employed analysis of variance (ANOVA) measures to determine if the teachers' SIOP training influenced the students' English

Table 3 New Jersey SIOP: Comparison vs. SIOP treatment group IPT mean scores

	03–04 Baseline			04–05 Year 1			05–06 Year 2			Change 03–04 to 05–06
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	
Oral proficiency level:										
Comparison	(192)	3.66	1.28	(169)	3.65	1.32	(168)	3.66	1.33	0.00
SIOP	(387)	3.67	1.37	(278)	3.76	1.26	(268)	4.00**	1.06	+0.33
Reading proficiency level:										
Comparison	(188)	3.95	.92	(169)	3.98	1.01	(168)	3.97	.92	+0.02
SIOP	(387)	3.82	.92	(278)	3.91	.88	(268)	4.10	.83	+0.28
Writing proficiency level:										
Comparison	(176)	4.16	.94	(169)	4.04	1.16	(168)	4.02	1.13	–0.14
SIOP	(386)	4.06	1.08	(278)	4.16	.94	(267)	4.32**	.83	+0.26
Total proficiency level:										
Comparison	(193)	3.69*	1.04	(166)	3.61	1.23	(168)	3.65	1.12	–0.04
SIOP	(386)	3.11	1.06	(278)	3.70	1.00	(267)	3.88**	.92	+0.77

Notes: * statistically significant in favor of Comparison group ($p < .05$); ** statistically significant in favor of Treatment group ($p < .05$)

language achievement. We focused on Year 2 because by then most of the treatment teachers were adequately trained in the SIOP Model and thus the SIOP intervention might have a meaningful impact on student achievement. We also calculated the effect sizes (Cohen's *d*) for the differences in results between the treatment and comparison districts.

Table 3 illustrates the comparison and treatment groups' average IPT mean scores for oral language, reading, writing, and total proficiency. Regarding oral language proficiency, the two districts were at about the same level in the baseline year, but treatment students overtook comparison students in Year 1 and continued to outperform them in Year 2. By Year 2, the average mean score in the treatment district was significantly higher than in the comparison district, $F(1, 434) = 8.49$, $p = .004$, $d = 0.29$. Reading had a similar trend except that treatment students overtook comparison students only in Year 2 and the differences in average mean scores did not reach statistical significance, $F(1, 434) = 2.49$, $p = .12$, $d = 0.16$. In writing, comparison ELLs had slightly higher performance in baseline year; however, in Years 1 and 2, treatment students had higher mean scores. By Year 2, this difference was statistically significantly higher, $F(1, 433) = 9.74$, $p = .002$, $d = 0.31$.

Total English proficiency level scores showed the same trends as the oral language and writing data results but were moderated by reading results. Nonetheless, we found that although comparison students had better total proficiency scores than treatment students in the baseline year, treatment students surpassed them in Years 1 and 2, showing a statistically significant difference in mean scores in Year 2, $F(1, 433) = 5.36$, $p = .02$, $d = 0.23$. Overall, the comparison scores were relatively flat or decreased over time while the treatment district scores increased for each measure.

The ANOVA results provided some evidence of SIOP as a predictor of achievement in oral language, writing, and total English proficiency. The analyses of β coefficients revealed that comparison students could expect to score lower than treatment students on all tests (Oral: β = –.32; Reading: β = –.12; Writing: β = –.29). For example, comparison students could expect to score about one third of a proficiency level lower on a scale of 0 to 5 than treatment students (Short et al., to appear).

The effect size calculations (interpretable as the standard deviation difference between the two groups) show the SIOP intervention had a small effect. The treatment group's scores were more than one fourth of a standard deviation higher than those of the comparison group for oral (0.29), almost one third of a standard deviation higher for writing (0.31), and close to one fourth for total English (0.23).

b Content area achievement: We collected and analysed student achievement data from New Jersey state tests in reading, math, social studies, and science for Grades 6–7[2]; reading, math, and science for Grade 8; and reading and math for Grade 11. The students in the treatment and comparison districts took these tests only once. The results showed a significant difference in mean scores in favor of SIOP students in the treatment district on six state content tests: in 2005, TerraNova reading ($p = .04$), language ($p = .03$), and total (reading + language + math) ($p = .02$) for Grade 6; and in 2006, NJ Ask language ($p = .01$) for Grade 6, NJ Ask language ($p = .01$) for Grade 7, and HSPA mathematics ($p < .01$) for Grade 11. There was a significant difference in mean scores in favor of students in the comparison district on one state content test: in 2005, TerraNova social studies ($p = .02$) for Grade 7. There were no significant differences between groups on the other 19 content tests (for more detailed results, see Center for Applied Linguistics, 2007).

The content achievement results indicate some promise for the SIOP Model but the number of student participants was very small for each test, and therefore the results are not generalizable. Further, because the state gives different tests in all three content areas for Grades 6, 7, 8, and 11, there are no pre- and post-test options for the same group of students. Moreover, with no tests administered in Grades 9, 10 or 12, students in Grades 8 and 11 do not have two successive years of testing in the same subject.

4 Discussion

As a result of this study, we felt the SIOP Model had merit as a successful intervention for academic literacy among ELLs. The significant differences in the average means in favor of the treatment group on oral language, writing, and total English scores indicate that the SIOP Model professional development had a positive impact on the development of English language proficiency among the English language learners in classes with SIOP-trained teachers. Although the IPT reading proficiency scores did not show significant differences between comparison and treatment groups, the trend favored the treatment schools.

The academic English scores for the treatment group improved significantly even though most of the ELLs' SIOP instruction took place in mathematics, science, and social studies classes. This suggests that the SIOP Model's attention to language development

influenced student English language performance and improved the quality of content area teaching. The comparison students' scores remained relatively constant over this time, representing what we more customarily find in districts with English language learners because each year new beginning-level students enter the ESL program while advanced level students exit. Therefore, the steady academic language growth of the treatment student group is striking.

There were small effect sizes on IPT oral, writing, and total English proficiency for ELLs with SIOP-trained teachers. When interpreting these differences, we should consider that effect sizes for treatment differences tend to be greatest at the primary grades with a steady decline as the grades progress (Bloom et al., 2008). Further, effect sizes for ELLs are often lower and more variable than those for native English-speaking students in literacy intervention studies (August & Shanahan, 2006). The English language learners in this study were in Grades 6–12 so, although small, the effect sizes of .29 to .31 for the difference in means on the oral language and writing measures show the potential of the SIOP training.

In response to our research question about teacher development, we found that 56% of the treatment teachers in Cohort 1 became high implementers of the SIOP Model after one year. Notably, 74% of the Cohort 2 teachers reached the high implementation level in just one year. We argue that the context of the SIOP Model initiative played a role in this difference. Cohort 1 teachers participated in a new, district-level initiative in Year 1. The coaching support was limited, the notion of focusing on language development in content courses was new, and a culture of working in a cross-disciplinary way was lacking. In contrast, Cohort 2 teachers entered an existing structure and joined a team of teachers and coaches who had already experienced success with the SIOP Model. The SIOP Model was also viewed favorably at that point by the administration, which devoted more staff time to coaching, thus affording teachers more support.

IV The CREATE SIOP experimental study

In the New Jersey SIOP study, we scaled up the initial CREDE SIOP research with a greater number of teachers, students, grade levels, and outcome measures. The next project expanded the program of research in four ways:

- We moved to an experimental design.
- We focused on middle-school science.
- We explored strategies to reduce the time needed for teachers to become high SIOP implementers; and
- We added native English speakers and former ELLs to the analysis.

The next investigation, 'The impact of the SIOP Model on middle-school science and language learning', was sponsored by the Center for Research on the Educational Achievement and Teaching of English Language Learners (CREATE) and funded by the US Department of Education. The multi-phase project began in the 2005–06 school year and was conducted by researchers at California State University Long Beach, the Center for Applied Linguistics, and the University of Houston.

Using a small cluster-randomized trial with randomization at the school level, the study investigated the impact of the SIOP Model on student achievement in middle-school science and tested alternative delivery systems of SIOP professional development. Science was selected because of its importance in schooling and because it was a recent addition to federal testing mandates. The research question was:

> What are the effects of the SIOP Model on the acquisition of academic language and science concepts among English language learners in middle-school science classrooms?

1 Phase 1: Pilot study

Phase 1 of the project was a pilot study to design and refine Life Science curriculum units in order to infuse SIOP Model instruction. The goal of the modified units was to jump start the teachers' SIOP implementation. The units were aligned to state standards in science, English language arts, and English language development. They were designed using the district textbook and other curriculum materials and followed the district pacing guide. The language objectives in the units targeted language and literacy skills necessary for ELL achievement in science. While lesson activities varied, key elements were present throughout: activities to practice all four language skills, frequent opportunities for student–student interaction, use of manipulatives and graphic organizers, modeling of lesson tasks, and review of key vocabulary and content concepts at each lesson's end. The units were revised for the Phase 2 study.

We also created and field-tested science language assessments aligned to the units to measure ELLs' acquisition of science vocabulary, reading, and writing skills. The process was informed by the World-class Instructional Design and Assessment (WIDA) *Access for ELLs*® test design and by research from the Center for Research on Evaluation, Standards, and Student Testing (CRESST) on writing test tasks. Specifically, these researcher-developed assessments included graphic support, text elaboration and simplification, items ranging in levels of difficulty, and task types similar to lesson activities. Each assessment had a reading passage on the topic being tested, such as photosynthesis, followed by multiple choice, short answer, and essay questions.

2 Phase 2: Randomized study

In Phase 2, 10 middle schools in one large urban district in southern California with large (over 25%) and moderate (4%–10%) numbers of ELLs were selected for this study. Schools in each category (large and moderate) were randomly assigned to either treatment (SIOP Model) or control (typical science instruction) conditions ensuring an equal distribution of ELL population type. Following randomization, but prior to data collection, two control schools withdrew, leaving a total of three control schools and five SIOP treatment schools.

Following state requirements, science teachers in both conditions had certifications or endorsements for teaching English language learners. However, treatment teachers ($n = 8$) were trained in the SIOP Model and provided with SIOP science curriculum units so that they would be prepared to implement the lesson plans effectively. Treatment teachers

received an intensive two-and-one-half-day workshop before the school year began where researchers introduced them to the SIOP Model components and second language acquisition theory. Participants watched videos illustrating effective classroom implementation of each component's features, rated the lessons using the protocol, justified their ratings, and engaged in a thorough group discussion of each feature. They also practiced SIOP techniques to deepen their understanding. Participants then reviewed the SIOP lesson plans in the Life Science units (cell structure and function, photosynthesis and respiration, cell division, genetics), including activity procedures, student handouts, and the assessments.

Life Science was taught at Grade 7 in this district and for only one semester. This situation unfortunately condensed the time for professional development, data collection, and potential impact. Over nine weeks, teachers taught the four units using the SIOP lesson plans. Coaching was provided to each treatment teacher by researchers experienced in implementing the model. The coaching process included three steps:

1. teacher and coach reviewed the lesson plan together in advance;
2. coach observed and rated the lesson using the SIOP protocol; and
3. the two held a debriefing and feedback session.

Control teachers ($n = 4$) taught the same four topics of study in the same time frame using their typical methods. They received neither SIOP training nor coaching. They were observed by researchers and their lessons were rated with the protocol as well.

Students in both conditions were given the CREATE science language assessments as a pre-test at the beginning of each unit and as a post-test at the end to measure growth in acquisition of science language. The essays were scored using the IMAGE writing rubric. At the end of the semester, all students were given a science content measure developed and administered by the district.

In the data analysis, we compared the assessment results of students in the SIOP classes ($n = 649$) to those of control students ($n = 372$). The sample included:

- students who were native English speakers (English Only = EO);
- students who had been redesignated as fluent English proficient for more than 3 (FEP3);
- students who had been more recently redesignated (FEP) (3 years or less); and
- English language learners.[3]

To determine if the SIOP instruction affected students' science language and concept development, hierarchical linear modeling (HLM) was used, with pre-test scores serving as the covariate. Our statistical model used students within sections, section within teacher, and school as the random effects. In addition, because student-level and teacher-level fixed-effect variables may influence student outcomes, we examined the students' pre-test science scores and their language classification (e.g. EO, ELL) as student variables, and condition (treatment or control) and level of SIOP implementation (high, medium, low) as teacher variables. The outcome variables were the composite post-test scores created by aggregating the four specific post-test assessments into separate scores for the essay and non-essay (multiple choice and short answer) components.

Table 4 CREATE SIOP: Science language mean scores for non-essay and essay components

	Control (n = 372)				Treatment (n = 649)			
	Pre-test		Post-test		Pre-test		Post-test	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Non-essay component	15.48	7.31	19.36	8.60	15.86	7.81	20.83	8.72
Essay component	41.77	20.89	56.41	24.62	43.78	24.72	61.21	29.63

Table 4 presents the mean scores for the composite non-essay and composite essay components. Results from the conditional ANCOVA model of HLM indicated that students in the treatment condition – regardless of language proficiency classification – outperformed, on average, those in the control, although not to a statistically significant degree. There was an approximate 0.9 point advantage ($\gamma = 0.9$, s.e. = 2.1, $t = .429$, $p = .67$) for students in SIOP schools on the non-essay component of the post-test and an approximate 5.5 point advantage ($\gamma = 5.5$, s.e. = 6.8, $t = .809$, $p = .418$) on the essay component.

We also considered the effect sizes using a pooled within-groups estimate of the standard deviation taken from a simple calculation of the within-group means and standard deviations (Hedges, 2007). The effect of SIOP instruction on the non-essay component of the post-test was associated with Hedges' $g = .103$, whereas the effect on the essay component of the post-test was $g = .197$. These results indicate small positive effects.

Based on our observations of the teachers in both groups, we were not surprised that significant differences between the treatment and control groups did not occur. Many SIOP-trained teachers did not implement the model to a high degree, and some control teachers implemented many SIOP features, resulting in high ratings on the SIOP protocol. For this reason we explored the relationship between teacher implementation level and student achievement, regardless of condition. The results indicated a positive relationship between teacher implementation level and average student gains. In other words, students whose teachers implemented the SIOP Model to a high degree performed significantly better on the assessments than students whose teachers were weak implementers ($R^2 = .22$, $p < .05$), emphasizing the importance of fidelity to the model (Echevarría et al., 2011).

3 Discussion

Several reasons may explain why the overall differences between the treatment and control groups were not stronger. First, at the research design level, power was limited due to the small number of schools and teachers. Although we had 27 sections in the SIOP group, 15 in the control group, and over 1,000 students in the study, we were limited with only eight schools and 12 teachers agreeing to participate. Ideally, the study would have involved a larger number of schools, more teachers, and a more balanced representation of sections within teachers.

However, increasing the sample size in any of these dimensions heightens the difficulty in implementing the SIOP treatment. Changing teacher practice requires significant time and on-going support (Saunders et al., 2009), and both are necessary for most teachers to reach a high level of implementation. In this study the intervention lasted only 9 weeks. It is unlikely that in this very short time teachers were able to develop a strong working knowledge of the SIOP Model. Moreover, this period of time was too brief for a change in instruction to significantly impact student achievement. Yet, the effect sizes, despite the short intervention, are affirmative, and it is possible that given more exposure to high quality SIOP science instruction, the students would have performed better.

The level of SIOP implementation varied greatly among teachers for many reasons including level of commitment to the project, skill, and teaching style. Although some variability in adoption of a new practice is expected, in a study such as this one, with a limited number of schools and teachers, this variability further reduces power and undermines the impact. The initial teacher professional development was only two-and-one-half days long and was conducted just prior to the beginning of data collection. It is possible that with longer, more sustained training, teachers would have implemented the SIOP features to a greater degree, which would have had greater effect on student achievement. Nonetheless, when teacher implementation was high, student achievement did increase.

V Challenges and benefits to scaling up SIOP research

It is expected in educational research for promising interventions to be scaled up. As we conducted this program of SIOP research over time, however, we realized that scaling up yields both challenges and benefits.

1 Teacher commitment

Teacher commitment can be a challenge or a benefit. In our CREDE SIOP study, we worked with a small group of dedicated, enthusiastic, middle-school teachers who were recommended for participation due to their status as excellent teachers of ELLs. Their commitment to the research and development of a model of sheltered instruction was strong. Their willingness to work collaboratively after school, to invite us in to observe in their classrooms, to regard videotapes of their instruction with a critical eye, and to test the organic model was beneficial all around. As we moved to the writing assessment study, they recruited additional teachers, most of whom also had a keen interest in the results.

In the New Jersey study, with more teachers and the high-school level in the mix, it took more time to persuade many treatment teachers that the research had value and that SIOP instruction was both doable and useful in their classrooms. Fortunately, we had two years to work with the staff and, as noted, the second cohort acculturated more rapidly and, as a group, learned the model faster. In this study, most of the participants were volunteers and we had strong administrative commitment. These factors added to the positive environment. At the comparison site, however, there was much less commitment to the research. Teachers tolerated our observations but showed scant interest in the investigation.

The CREATE study in contrast assigned schools and teachers randomly to treatment or control conditions. While it provided science lessons to jump-start SIOP implementation in class, the professional development time was considerably truncated. Commitment felt qualitatively different here. More teachers went through the motions and despite coaching sessions to support their efforts, interest in implementing the lessons well was decidedly lacking.

Teacher commitment correlated to fidelity of implementation across the three studies. In general, the more committed the teachers, the harder they tried, the more they sought out coaching guidance, and the more enthusiasm they had. Some still took two years to reach high levels of implementation as in the New Jersey study, but they made consistent progress during that time.

2 *Accountability pressures*

We believe that the US educational context also played a role in the teachers' commitment. We began the research program before the 'era of accountability' was the norm. Prior to the No Child Left Behind's widespread enactment in the 2002–03 school year, teachers were less focused on test scores and had relatively more freedom in their classrooms. Teachers felt they had more time for activities outside their regular duties, such as collaborative research projects. They had no pressure to raise test scores in order to ensure that a school would make mandated adequate yearly progress (AYP) targets.

Even when we began our New Jersey study in the summer of 2004, the pressure was less than we find today. Schools may not have been performing as well as desired, but since the time that the study began, severe sanctions, such as school closures, have been carried out based on school performance. This, along with several years of bad press and growing resentment of students who 'bring down' AYP have changed the US classroom climate. At the time of the study, however, the New Jersey study teachers were concerned about student achievement, but not overly so. Moreover, New Jersey had an alternate graduation assessment available to students who did not pass the state high-school tests, and this option eased the pressure for teachers and students.

The climate was different during the CREATE study. Teachers were stressed by test scores and school performance. More paperwork and more test-taking practice were asked of them. They felt that they had less time for non-essential activities, which is how they viewed research participation. The situation was further impaired because we had less time to get to know the teachers and to help them understand and implement the SIOP Model.

3 *Comparison/control groups*

The selection of the comparison district in our New Jersey study posed an unexpected challenge. We easily identified the treatment district and received approval from the district superintendent and the state Department of Education. However, we needed a separate comparison district because the middle schools were not matched (one was a designated Title I school, a low socioeconomic indicator) and high-school students and teachers could not be placed in distinct groups. We looked for a district that matched on

several factors, i.e. linguistic diversity, socioeconomic status, achievement scores on state exams, and an ESL language program design; but convincing a district to act as the comparison site was difficult. Districts declined for several reasons. Some with predominantly native Spanish speakers provided a bilingual program in the middle schools. A few had participated in another intervention in recent years and were reluctant to take on a new project. Some districts objected to being labeled a comparison site, despite promised anonymity and free professional development. As a result, we caution other researchers to anticipate the challenge of securing a comparison or control site. As an educational community, we may need to develop new types of incentives for such participation.

4 Professional development

The professional development settings and time frames seemed to make a distinct difference in the outcomes of these studies. Having sufficient time to learn about the SIOP Model, try it out in class, and get feedback from coaches led to a higher level of implementation and more statistically significant results on measures of student academic literacy. In our CREATE study, we tried to condense the time needed to learn the model in order to test whether it could be shortened with the addition of SIOP-infused units and to conform to the district calendar. We found that a few days of training prior to data collection, even with lesson plans and coaching, were insufficient for teacher fidelity and significant student performance. However, in this study, the trend favored the treatment group and there was a small effect size, signaling some room for optimism if the professional development time were increased and if data collection occurred after teachers were more proficient in the model.

VI Suggestions for future research

In the New Jersey and CREATE studies, teachers were still engaged in the SIOP professional development process while data collection occurred. Some teachers had not completed the training and some did not implement the SIOP Model to a high degree before student achievement was assessed. Further, some students only had teachers who were low implementers of the model. Given the complexity of the model, should a similar study be done in the future, we recommend more time be given for teacher implementation before collecting student achievement data and an analysis of student performance linked to teacher level of implementation be conducted. We were unable to do this analysis in the New Jersey study because the subset of students with solely high or low implementers was too small for comparison purposes. The analysis in the CREATE study showed, however, that teacher level of implementation matters for student performance.

We also recommend a future study collect and analyse student performance data on content area exams, with pre- and post-measures, so that the impact of SIOP instruction on content achievement could be investigated. The content tests in New Jersey, as in most states, did not fit this goal. Students do not take the same test two years in a row. To examine student achievement then, other content tests are needed, but in the current educational climate in the USA with heavy testing already in place, more assessments are a burden that teachers and students may not deserve to bear.

A longitudinal, randomized study of the SIOP Model would also be welcome, but student mobility would need to be accommodated because turnover is high in these urban settings. Ideally, a randomized experimental study could look at the effects of SIOP Model instruction on the same cohort of students over three to five years and compare their achievement to that of a control group of students. Some schools are conducting program evaluations along these lines, but they usually lack a comparison or control group.

VII Conclusions

It is not often that educational researchers have a chance to build, test, and refine a model over an extended time period. We have had the advantage of working on SIOP research and professional development for more than 10 years. This has given us a chance to test the model in different settings and subject areas, with quasi- experimental and random experimental designs. We have studied ways to enhance the professional development and focused on one subject area to see if such concentration of teachers and curriculum units might have an impact. We have seen positive effects for ELLs, former ELLs, and native English speakers.

The findings from these studies indicate that the SIOP Model offers a promising approach to professional development that would improve the quality of instruction to ELLs and enhance their English language achievement. Although the effect sizes were small, they represent positive results. So far, we have found that teachers with SIOP Model training need 1–2 years of support to become high implementers. We will continue to test methods for reducing the time, but we acknowledge that changing teacher practice is a long-term endeavor. School reform policies need to anticipate and plan for such time frames and ought to provide not only a series of workshops but also additional support through coaching, lesson planning, and other technical assistance.

Acknowledgments

We would like to acknowledge Carolyn Fidelman and Mohammed Louguit of the Center for Applied Linguistics, Washington, DC, Kristin Powers of California State University Long Beach, CA, and David Francis of the University of Houston, TX for their assistance with statistical analyses, and Thomas Bauder, Beverly Boyson, Rebecca Canges, Lindsey Hillyard, Jennifer Himmel, Justine Hudec, Christopher Montone, and Marsha Vargas for their research assistance during these studies. This work was supported by the US Department of Education's Office of Educational Research and Improvement (grant number R306A60001), Institute for Education Sciences (grant number R305A05056), the Carnegie Corporation of New York (grant number 7609), and the Rockefeller Foundation (grant numbers 2004WC144 and 2005WC144).

Notes

1 Although the state of New Jersey switched language proficiency tests in the second year of the study, our two districts gave the IPT again.

2 New Jersey changed tests during the study. In the first year, Grade 6 and 7 students took the TerraNova in reading, language, math, social studies and science. In the second year, they were only tested in reading and math on the NJ Ask.

3 Because these were heterogeneous science classes, not only ELL, our research question was broadened to include all students.

References

August, D. & Shanahan, T. (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth*. Mahwah, NJ: Erlbaum.

Bloom, H., Hill, C., Black, A., & Lipsey, M. (2008). *Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions*. MDRC Working Papers on Research Methodology. New York: MDRC.

Center for Applied Linguistics (2007). *Academic literacy through sheltered instruction for secondary English language learners.* Final Report to the Carnegie Corporation of New York. Washington, DC: Center for Applied Linguistics.

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

Crandall, J. (1993). Content-centered learning in the United States. *Annual Review of Applied Linguistics*, *13*, 111–26.

Echevarría, J., Richards-Tutor, C., Chinn, V., & Ratleff, P. (2011). Did they get it? The role of fidelity in teaching English learners. *Journal of Adolescent and Adult Literacy*, *54*, 425–34.

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

Echevarría, J., Vogt, M.E., & Short, D. (2008). *Making content comprehensible for English language learners: The SIOP Model*. 3rd edition. Boston, MA: Pearson/Allyn & Bacon.

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2006). *Educating English language learners: A synthesis of research evidence*. New York: Cambridge University Press.

Guarino, A.J., Echevarría, J., Short, D., Schick, J., Forbes, S., & Rueda, R. (2001). The sheltered instruction observation protocol. *Journal of Research in Education*, *11*, 138–40.

Hedges, L.V. (2007). Effect size estimation in cluster-randomized trials. *Journal of Educational and Behavioral Statistics*, *32*, 341–70.

Hudec, J. & Short, D. (Prods.) (2002a). *Helping English learners succeed: An overview of the SIOP Model* [Video]. Washington, DC: Center for Applied Linguistics.

Hudec, J. & Short, D. (Prods.) (2002b). *The SIOP Model: Sheltered instruction for academic achievement* [Video]. Washington, DC: Center for Applied Linguistics.

Illinois State Board of Education, Assessment Division (2004). *The Illinois State Assessment: Technical Manual 2004*. Retrieved from: http://www.isbe.net/assessment/pdfs/isat_tech_2004.pdf (March 2011).

Saunders, W.M., Goldenberg, C.N., & Gallimore, R. (2009). Increasing achievement by focusing grade level teams on improving classroom learning: A prospective, quasi-experimental study of Title 1 schools. *American Educational Research Journal*, *46*, 1006–33.

Short, D. (2006). Content teaching and learning and language. In K. Brown (Ed.), *Encyclopedia of language and linguistics, volume 3* (pp. 101–05). 2nd edition. Oxford: Elsevier.

Short, D., Fidelman, C., & Louguit, M. (to appear). Developing academic language in English language learners through sheltered instruction. *TESOL Quarterly*.

Short, D., Hudec, J. & Echevarría, J. (2002). *Using the SIOP Model: Professional development manual for sheltered instruction*. Washington, DC: Center for Applied Linguistics.