DOES SIOP RESEARCH SUPPORT SIOP CLAIMS?

Stephen Krashen

International Journal of Foreign Language Teaching. 8,1: 11-24, 2013.

"The SIOP Model* is a research-based and validated model of sheltered instruction. Professional development in the SIOP Model helps teachers plan and deliver lessons that allow English learners to acquire academic knowledge as they develop English language proficiency." (http://www.cal.org/siop/)

"Based on years of research, the Sheltered Instruction Observation Protocol (SIOP®) Model is a scientifically-validated framework for improving the academic achievement of English learners." (http://siop.pearson.com/events-training/siop-training-for-teachers-virtual-institute.html)

The goal of this article is to determine whether these claims are supported.

SIOP stands for "Sheltered Instruction Observation Protocol" and is a rubric, a checklist to see if teachers are following certain procedures while teaching intermediate second language acquirers. It is also the foundation of a method for doing content-based second language teaching.

THE SIOP CHECK-LIST: A MIXED BAG

The field of language education is now engaged in a struggle between two hypotheses that present fundamentally different views about how language is acquired. SIOP ignores this struggle, and simply accepts both as valid.

The "Skill-Building" Hypothesis assumes that students first need to consciously learn their "skills" (grammar, vocabulary, spelling), and that only after skills are mastered can they actually use these skills in real situations; they are made "automatic" as students use them in writing and speaking, and students can fine-tune their rules when they are corrected. Skill-Building thus depends on conscious learning, output, and correction.

In contrast, the Comprehension Hypothesis claims that we acquire language and develop literacy in only one way: when we understand messages; that is, when we understand what we hear and what we read. We do not acquire language when we produce it, either in speech or writing, and we do not acquire language when we study grammar rules or memorize lists of vocabulary.

According to the Comprehension Hypothesis, language acquisition occurs subconsciously; while it is happening, we do not know it is happening. We think we are having a conversation, listening to the radio, watching TV, watching a movie, reading a book (which of course we are), but at the same time, without realizing it, we may be

acquiring language. Also, after we have acquired some aspect of language, we are generally not aware of it. The new knowledge is stored subconsciously in our brains.

According to the Comprehension Hypothesis, our knowledge of grammar and vocabulary emerges as a result of getting comprehensible input. According to the Skill-Building Hypothesis, we first learn grammar and vocabulary, but can only use them after they are made automatic.

To provide a bit more detail, the Comprehension Hypothesis claims that we acquire language when we understand messages containing aspects of language that we have not acquired, but are developmentally ready to acquire. We achieve comprehension through context, the use of our current linguistic competence and background knowledge.

There is overwhelming evidence that the Comprehension Hypothesis is correct; nearly all of our competence in language and literacy comes from understanding what we hear and read:

- Students in beginning-level second language classes based on the Comprehension Hypothesis consistently outperform students in classes based on skill-building on tests of communication, and do at least as well as, and often slightly better than, students in skills-based classes on tests of grammar (Krashen, 2003).
- At intermediate levels, comprehensible input is delivered in two ways:
- One is the use of "sheltered" subject matter teaching, classes in which students study subject matter made comprehensible in the second language. The focus of these classes is subject matter, not language, i.e. students are not tested on language but on subject matter. Students in sheltered subject matter classes acquire as much language as or more language than students in traditional classes and also learn impressive amounts of subject matter (Krashen, 1991; Dupuy, 2000). (1)
- A second source of academic comprehensible input for intermediates is reading, especially self-selected free voluntary reading. A variety of studies have confirmed the effectiveness of free voluntary reading in both first and second language development: (1) correlational studies show that those who read more develop higher levels of vocabulary, reading, and writing ability; (2) studies confirm that students who participate in self-selected reading programs in school (sustained silent reading) do as well as or better than students not in these programs in reading and vocabulary development; (3) case histories of those who attained high levels of literacy and language through self-selected reading, without significant amounts of study, confirm the power of reading. (Krashen, 2004).

There are also independent reasons to reject Skill-Building as a significant source of language competence. (1) The systems to be consciously learned are extremely complex. For example, the system of grammar of all languages has not even been fully described by linguists. Competent first and second language users typically know far more vocabulary than can possibly be learned one word at a time. (2) The necessary components of Skill-Building, error correction and output, are not frequent enough to

account for more than a small percentage of our competence in language (Krashen, 1994).

In addition, studies confirm that error correction has only a limited impact on language proficiency, and that increasing output does not increase language proficiency (Truscott, 1996; 1999; Krashen, 1994).

Consciously learned knowledge of language can contribute to language performance, but it is highly limited. Very severe conditions must be met; the user must know the rule, be thinking about form, and have time to apply the rule. The second two conditions are usually met only on grammar tests and when we have time to edit our writing (Krashen, 2003).

An important corollary of the Comprehension Hypothesis is that producing language per se does not cause language acquisition; talking and writing are not "practicing." Rather, the ability to speak and write emerges as a result of language acquisition, of receiving comprehensible input. "Forcing" speech before the acquirer is ready will not help acquisition, and often only results in tension (Krashen, 2007).

SIOP considers the Comprehension Hypothesis and the Skill-Building Hypothesis to be equal partners in developing second language competence.

SIOP and Sheltered Subject-Matter Teaching

SIOP embraces the core principles of sheltered subject-matter teaching:

Two items—Items 7 and 8—of the SIOP rubric call for teachers to provide background knowledge, a crucial aspect of making input comprehensible. They call for teachers to explicitly link concepts to students' background knowledge and link past learning and new concepts.

Three items—10, 11 and 12—are clearly dedicated to making input more comprehensible; they call for teachers to use clear and less complex speech, provide clear explanations, and use techniques such as modeling and using visuals to provide context.

SIOP and Skill-Building: Output Emphasis

SIOP gives input and output the same importance in items 6 and 22: "Meaningful activities that integrate lesson concepts … with language practice opportunities for reading, writing, listening, and/or speaking," and "Use activities that integrate all language skills (i.e. reading, writing, listening, speaking)."

Moreover, in item 16, SIOP encourages "elaborated responses," which may force students to attempt to produce what they have not yet acquired.

SIOP and Skill-Building: Focus on form

In addition to content objectives, SIOP wants us to have "clearly defined language objectives" (items 2 and 24). SIOP's focus on form emphasis is confirmed by item 21, which requires activities to "apply content *and language knowledge* (italics mine) in the classroom"; in other words, to consciously monitor newly learned (not acquired) language, a difficult task that does not result in language acquisition.

SIOP and Skill-Building: Correction

SIOP requires correction of output in item 29: "regularly provides feedback to students on their output," including language output.

SIOP and Skill-Building: Vocabulary Teaching

SIOP calls for an emphasis on "key vocabulary," which appears to be a focus on learning new vocabulary (item 9), especially since it requires new vocabulary not only to be "introduced" but also to be "written, repeated, and highlighted …". In addition, item 27 calls for a "comprehensive review of key vocabulary."

In summary, items of the SIOP ask teachers to base their approach on the Comprehension Hypothesis (items 7, 8, 10, 11, 12) and six items ask teachers to base their approach to a large extent on the Skill-Building Hypothesis (items 2, 6, 9, 16, 22, 24, 27). (For more description of the SIOP as well as the entire protocol, see Echevarria, Short, and Powers, 2006).

Thus, research testing the impact of SIOP as a whole will not tell us the basis for its success or failure. No study has examined the impact of SIOP that is sensitive to teachers teaching in agreement with the Comprehension Hypothesis-based items versus the Skill-Building Hypothesis-based items.

One would predict that SIOP teachers with higher CI scores would stimulate more language acquisition, while those with higher SB scores would stimulate more conscious language learning. Unfortunately, in the studies that I describe below, the distinction in different aspects of SIOP are ignored: Both the SIOP rubric and SIOP training are treated as a whole, with no analysis of its parts.

THE VALIDITY STUDY

Guarino, Echevarria, Short, Schick, Forbes and Rueda (2001) claims to be an evaluation of the reliability and validity of the SIOP. Four raters analyzed six videos of lessons, three of the videos "deemed by specialists to be highly representative of the tenets of SI" (sheltered instruction), while the other three were not. Guarino et al. reported that the experts gave the "highly representative" lessons higher ratings on the SIOP than they gave the unrepresentative lessons. This is confirmed in table 1 (from Guarino et al.'s

table 2), which gives separate results for each of three components of the SIOP (Preparation, Instruction, Review/Evaluation).

Note that for some components, the mean score for the SIOP representative classes fell well short of a perfect score, especially for Review/Evaluation, but it is clear that the SIOP ratings assigned to SIOP representative lessons are higher than those assigned to the NON SIOP classes.

Table 1: Data from the Validity Study

Subcomponent	items	Maximum	SIOP lessons	NON SIOP
Preparation	6	42	30.6 (4.05)	13.4 (3.8)
mean item		7	5	2.2
Instruction	20	140	129.1 (19.03)	54.02 (14.6)
mean item		7	6.5	2.7
Review/Evaluation	4	28	15.6 (4.5)	5.9 (2.9)
mean item		7	3.9	1.5
Total	30	210	175.3	73.3
mean item		7	5.8	2.4

From Guarino et al., 2001, table 2

THE FIDELITY STUDY

Echevarria, Richards-Tutor, Chinn, and Ratleff (2011) claim to show that higher scores on the SIOP are related to greater gains by students; in other words, they claim that the SIOP has predictive validity. A close look at their findings reveals several serious problems.

SIOP ratings

In this study, a group of 12 teachers were rated on the 30 features of the SIOP. Echevarria et al. state that each teacher was observed five times, approximately every other week (p. 430). Each teacher's score was an average of the five observations. Eight teachers had been trained on SIOP and four had not, but SIOP scores for all 12 teachers were used in the analysis.

The measures

A pre-test and a post-test was used, and students' scores were the difference between them (gain scores). The tests asked students to read a paragraph about science, based on material taught in class, and answer multiple-choice and fill-in questions. Means and standard deviations for the test were not presented. Gain scores and SIOP scores can only be crudely approximated from Echevarria et al.'s figure 3.

The students

We are told very little about the students, only that they attended seventh grade in a "large urban school district with high numbers of ELs" (p. 428) and that there was a total of 1021 students, 649 taught by teachers trained in SIOP and 372 taught by teachers without SIOP training. We do not know how many of the students were ELLs.

The Statistical Analysis

The graph presented by Echevarria et al. does indeed show a positive trend, and contains the information r2 = .2183. Apparently, Echevarria et al. performed a simple regression analysis. If so, this means that knowing a teacher's SIOP score gives us about 22% of the information we need to predict the gain their students made, a modest result.

An r2 of .2183 for a simple regression analysis is equivalent to a correlation (r) of .46. For a sample size of 12, this size correlation falls just short of statistical significance (one-tail test). This is not mentioned in Echevarria et al., nor is there any discussion of statistical significance.

It also needs to be pointed out that a correlation using a sample size of only 12 is not "powerful" enough to detect significant differences.

"Statistical power" is a number that tells us what the chances are of detecting a difference if one exits.

For n = 12 and r = .46, assuming a one-tail test, and following Welkowitz, Ewen and Cohen (1982), pp. 226-227, I calculated a power level of .44 (one-tail test). This is well below Cohen's recommended minimum level of power of .80 (Cohen, 1977), but few published studies achieve this level of power (Jennions and Mollerb, 2003). This flaw, however, should be noted and claims made for predictive validity made more tentative. To achieve the .8 level, SIOP would need a sample size of at least 30 subjects, according to my calculations.

Summary: The Fidelity Study

A great deal is missing from the "fidelity study," including means and standard deviations for all measures, and details about the students. Moreover, as discussed earlier, the SIOP itself consists of contradictory features representing different views of language acquisition; we do not know which features were responsible for the results, if the results are indeed valid.

Finally, the results themselves are modest, and fall short of statistical significance. The test run lacked sufficient power to detect a difference if one really existed.

The data presented in Echevarria et al. clearly do not support their strong conclusion: "As our study shows, there is a direct relationship between level of implementation and

student achievement" (p. 433). At best, the data as presented suggest that there might be such a relationship. (2)

COMPARING SIOP

A small number of studies have been done in which the performance of classes of SIOP-trained teachers are compared to the performance of comparison classes, those not taught by SIOP-trained teachers. As discussed earlier, because the SIOP represents competing hypotheses about language acquisition, no matter how this research comes out, we don't know what is causing the effect, if any. In addition, as we will see, comparison groups are typically poorly described in this research. In other words, such comparisons are not useful theoretically.

As noted earlier, it has claimed that SIOP is supported by scientific research. As we will see, this is an exaggeration. Inspection of the individual studies reveals that the effects are quite weak.

I review here the only four studies I know of in which a group of students taught using SIOP methods is compared with other students. Three of the four studies were done by the SIOP developers.

In addition to the usual method of reporting results (significance levels), I present effect sizes, the best way to compare studies (table 2). Effect sizes tell us the impact of a treatment, to what extent one group was better than another.

Table 2: Comparison studies of SIOP

Study	Measure	effect size	significance level	n: SIOP/comp
Echevarria et al 2006	WRITING			
	language production	.32/.26 (1)	0.026	
	Focus	.38/.24	0.055	
	support/elaboration	.34/.24	.103(2) ns	
	Organization	.52/.29	0.018	
	Mechanics	.26/.21	0.044	
	Total writing	.47/.35	0.001	238/77
Echevarria, Richards-Tutor, Canges,& Francis, 2011	SCIENCE			649/372
	(1) multiple choice, short answer	0.102(3)	.673 ns	
	(2) essay	0.197	.418 ns	
Short et al 2012	Writing	0.31(4)	.001 (5)	267/168
	Oral	0.29	.002	
	Reading	0.16	.06	
McIntyre et al 2010	Reading	0.16(4)	> .05 ns	50/57

(1) Effect sizes in Echevarria et al. (2006) were calculated by SK in two ways: Following Morris (2008), which takes pretest scores into consideration, (2) using adjusted posttest scores and the pooled posttest standard deviation. Echevarria et al. (2006) reported a total effect size of .83, but this was based only on the gain made by the SIOP group without consideration of the comparison group gain.
(2) significance level was incorrectly reported in the published paper as 1.03.
(3) effect size = Hedges g, reported by the authors, for all students combined. For English learners only (n = 105, 112), I calculated effects sizes of .062 for the nonessay test and .087 for the essay test, based on adjusted means and standard deviations derived from the standard errors in Echevarria et al., table 6.
(4) effect size for Short et al. calculated by SK following Morris (2008).
(5) significance level recalculated by SK, and converted to one-tail values. One-tail is required here, as the investigators predicted the direction of the effect. Other p-values in this table were reported by the authors; the authors did not state whether they were one-tail or two-tailed, and recalculation was not possible from the information given.

The effect sizes reported in table 2 are all positive, which means that the SIOP group did better than the comparison group in all cases. But the effect sizes are quite modest, and not always statistically significant.

Interpreting Effect Sizes

Cohen (1977) presents the following interpretations of effect sizes:

Small effect size: d = .2. This is equivalent to a correlation (r) = .1 and r2 = .01, which means that it accounts for about 1% of the variability; if the squared correlation between two variables is r2 = .01, knowing the value of one variable is 1% of what we need to predict the value of the other variable.
Medium effect size: d = .5. "A medium effect size is conceived as one large enough to be visible to the naked eye" (Cohen, p. 26). d = .5 corresponds to r = .24, and r2 = about .06, meaning that knowledge of the value of one variable is 6% of the information needed to accurately predict the value of the second variable.
Large effect size: d = .8, which is equivalent to r = .37 and r2 = .14. This difference is "grossly perceptible" (Cohen, p.27).

In table 2, most of the effect sizes are small. In fact, the only effect size to reach the medium effect size level was one component of the writing measure in Echevarria et al. (2006), using one method of computing effect sizes, and this occurred because the comparison group made no gain at all. The high score for this one component influenced the combined score.

I calculated the average (mean) effect size for all studies comparing SIOP and comparison groups in two ways, both producing similar results (table 3), with the average mean effect closer to "small" than "modest" according to Cohen's criteria.

Table 3: Mean effect size for all SIOP comparison studies
1. Unweighted for sample size, using the average value from each study:
Echevarria et al. 2006: .47
Eechevarria et al. 2011: .15
Short et al. 2012: .25
McIntyre et al. 2010: .16
mean = .26
2. Weighted by sample size (of experimental group), using the average value from each study = .33.

Note: For Echevarria et al., 2006, only effect sizes calculated according to Morris' method (Morris, 2006) were used for the calculations in table 3. Using the second method, based on adjusted post-test scores, gives an unweighted mean of .25 and a weighted mean effect size of .21.

In addition to these unspectacular results, further inspection of table 2 shows that the results were not statistically significant in two of the four studies.

At best, SIOP studies show a very modest superiority to comparison groups. Taken at face value, without consideration of the details of the studies, the results clearly do not support the frequently-repeated claim of SIOP authors that SIOP has been demonstrated to be effective by scientific studies. There are, in fact, only four studies, three done by SIOP authors. The effect sizes are very modest—and often small—and two of the four studies show non-significant results.

But even these results may be exaggerated.

Comparison groups

As presented in table 4, there is little information about the comparison groups. We do not know to what extent the comparison groups differed from the SIOP groups and whether they had more or less comprehension-based or skill-based instruction. This is crucial. Did the "sheltered" classes in the comparison groups in Echevarria et al. (2006) conform to the comprehensible input-based items in the SIOP? Were they similar to sheltered classes found to be superior to traditional instruction in Krashen (1991) and Dupuy (2000)?

Table 4: Description of comparison group treatment

Echevarria et al 2006	"sheltered"; no other details
Echevarria et al 2011	used "methods they would "normally use" (p. 338)
Short et all 2012	no description
McIntyre et al 2011	no description

In Echevarria et al. (2011), two comparison schools dropped out of the project.

Description of Students

SIOP is intended for speakers of English as a second language, but the authors suggest that it will be useful for all students. It is not clear in all the studies what percentage of ELLs was included. Nor is SES information supplied in all studies.

Table 5: Description of subjects

Echevarria et al 2006	Students	% ELL
	grades 6-8; low SES	100%*
Echevarria et al 2011	grade 7	all levels ; 30% ELL
Short et al 2012	grades 6 to 12; low SES	mixed
MeIntyre et al 2011	"elementary"	100%

* some native speakers of English were in some classes, but "were not part of the data collection" (p. 262).

Vital information is lacking on the subjects in nearly all cases (table 5). In McIntyre, we are told only that students were at the elementary level and were ELLs.

SIOP teachers: a special group?

As presented in table 6, teachers in the experimental/SIOP group were quite different from those in the comparison groups. All had extensive SIOP training, which in itself is not a surprise and could be considered as evidence for the ecological validity of these studies. But they were also special in that they were undoubtedly enthusiastic about their teaching method. In one study (Echevarria et al., 2006), SIOP teachers were "nominated" for the study, in another (Short et al., 2012), they were volunteers.

Table 6: SIOP teachers: training and selection

Study	SIOP training	Selection
Echevarria et al 2006	1 to 2 years, then "reunion meetings"	"nominated"
Echevarria et al 2011	2.5 days, feedback every other week	
Short et al 2012	7 days, + daily coaching	most volunteers
McIntyre et al 2011	50 hours	only those with perfect SIOP scores

There are unknown factors in several studies (nature of the comparison group, percentage of ELLs in classes, SES of students) and known factors that quite likely favored the experimental classes (specially selected or volunteer teachers). These problems make the empirical results even less impressive. (3)

Conclusions: Comparison Studies

There are few studies in which SIOP is compared to competing methods, and we are not sure what those competing methods were. There also exists an unusual number of flaws and gaps in the studies, and results are only modest, despite a large investment in SIOP training. Even if SIOP were shown to be successful, because SIOP is a mixed bag, we would not know what is responsible for the results because the effect of different parts of SIOP was not separately analyzed.

Additional Studies

A number of additional studies have involved SIOP.

Honigsfeld and Cohen (2008) is cited in Echevarria, Richards-Tutor, Canges, and Francis (2011), p. 337, as one of a series of articles showing that SIOP has "been shown to improve student achievement."

Honigsfeld and Cohen is a report of 22 teachers who participated in a graduate education course that covered "Japanese lesson study" and SIOP. The class included lesson planning using SIOP, which teachers implemented in their own classes. Honigsfeld and Cohen reported that the teachers were successful in applying SIOP and a year later the

teachers said that they still used it to at least some extent in their classes. There were no tests, and no questionnaires or formal interviews of the teachers. Honigsfeld and Cohen's report contains no clear evidence showing that SIOP was "shown to improve student achievement."

In Batt (2010), 15 teachers had a summer institute focusing on SIOP or had attended a "national SIOP institute" of unknown duration. The teachers were elementary school teachers, were "purposefully selected" and were considered "teacher leaders" in their schools. Betts provides little description of their students, only that the schools they attended were high poverty.

Teachers gave the initial training high evaluations, considering it to be effective, and 12 out of the 15 of the teachers said they were highly committed to implementing SIOP in their classrooms. After the training, eight of the 15 said they implemented the model in their classes "to a great extent." Apparently, 12 teachers talked the talk, and eight actually walked the walk.

The teachers then had a year of coaching and more workshops, narrowly focused on aspects of SIOP. All of the teachers said that they implemented the SIOP model "to a great extent" during or after the coaching.

Batt gives the impression that there was joyous participation and compliance by all of the teachers all of the time, and that the teachers were happy to work overtime (at one school, they met with their coach "before or after school, during lunch …") and that they wanted even more time with the coach. Another interpretation is that it took a summer institute and a whole year of coaching to bring these teachers into line.

The only evidence that SIOP improved student performance after the coaching is that the teachers felt that their students had improved. All teachers agreed or strongly agreed with the statement, "I have seen improvement in student achievement as a result of using the SIOP model."

In Whittier and Robinson (2007), 7th and 8th graders, ELLs from a high poverty area, participated in game-like activities to teach them the principles of evolution in ten 60-minute classes over one month.

The class was described as "sheltered" and was taught in agreement with "parts of" SIOP, but we are not told which parts. No SIOP scores of the teachers were presented. In fact, no information about the teacher or teachers was presented.

Whitter and Robinson reported that students improved in knowledge of the principles of evolution on a test from their science textbook, gaining from 26.9% correct to 42.3% correct, a gain of 15.4%. Apparently the same test was given as a pre and posttest.

We are not told how many items were on the test. If there were 30 items, the students improved from about 8 correct to about 12.5, or 4-and-a-half items, a very modest gain considering the time invested.

Students were also asked to write about a topic studied in the unit, apparently about the difference between a specialist and a generalist, both as a pre and posttest. They improved in ratings from 2.1 to 3 out of four, also a modest gain.

There was no comparison group, and there was no statistical analysis, even though this study was done by science teachers. All we can conclude from this study is that students learned a little about evolution, and it is not clear what aspects of SIOP were used.

SUMMARY AND CONCLUSIONS

1. The SIOP rubric supports opposing views of how language is acquired and how literacy is developed. Five SIOP items out of 30 are consistent with the Comprehension Hypothesis. Another six are consistent with the Skill-Building Hypothesis. These two hypotheses are not complementary; they are rival hypotheses.
2. The measure of SIOP validity is based only on the judgments of four experts who observed six lessons, three considered to be consistent with items on the SIOP rubric and three inconsistent.
3. The claim for SIOP predictive validity is based on data from only 12 teachers. The strength of the prediction is modest and did not reach statistical significance, r2 = .22, which means that knowing SIOP scores provides 22% of what is needed to predict student progress. Also, the sample size is too small for confidence in the conclusions.
4. Studies comparing SIOP-trained teachers with non-SIOP trained teachers provide only marginal evidence for SIOP's superiority. Effect sizes are generally small, there are only four studies (three done by SIOP creators), and in two of the four studies, the differences are not statistically significant. In addition, important information about comparison groups and the students taught is often missing, and it appears that SIOP teachers were specially selected in several studies.

Any of these four observations is enough to raise serious doubts about SIOP. All four should produce profound skepticism about SIOPs claims that it is a "research-based and validated model."

Because of the widespread use of the SIOP and its far-reaching advertising, published research supporting the SIOP should be made of sterner stuff. (4)

Acknowledgment: I thank Jim Crawford and Sharon Adelman Reyes for help and insightful discussion.

NOTES

1. Krashen (1991) and Dupuy (2000) are reviews of published studies comparing various kinds of sheltered classes to traditional methods. These articles and the many studies they cite were apparently missed by the Pearson Publishing Company, who claims that SIOP is "the only scientifically validated model of sheltered instruction." (http://www.pearsonschool.com/index.cfm?locator=PS1sAt). Not only have these studies validated sheltered instruction, but as we shall see later, SIOP lacks convincing scientific validation.
2. From the description of the teachers, students and measures, it is likely that the data from this study were taken from another SIOP study, Echevarria, Richards-Tutor, Canges, and Francis (2011), which appears to contain some of the information missing from the fidelity study. Echevarria, Richards-Tutor, Chinn, and Ratleff (2011), the fidelity study, does not cite Echevarria, Richards-Tutor, Canges, and Francis (2011).
3. In Echevarria et al., 2006, "the researchers (authors and their research assistants) administered the test and monitored the classrooms during testing" (p. 204). Since the authors have a professional and financial interest in the outcome, it would have been better to avoid this potential confound.
4. Studies of sheltered subject matter show that students in sheltered classes learn an impressive amount of subject matter, sometimes as much as native speaker comparison students (Krashen, 1991, Dupuy, 2000). There is little data on subject matter learning among SIOP students. Short, Echevarria, and Richards-Tutor (2011), reporting on data published later as Short, Fidelman and Louguit (2012), reported that SIOP trained students did significantly better than comparisons on six New Jersey state content tests, comparisons were better on one, and there was no difference on 19 other tests (Short et al. provided only p-levels). Short et al. point out that "the content achievement results indicate some promise for the SIOP model but the number of student participants was very small …" (p. 371) and pre-post test comparisons were not possible.

REFERENCES

Batt, E. 2010. Cognitive coaching: A critical phase in professional development to implement sheltered instruction. Teaching and Teacher Education: An International Journal of Research and Studies, 26(4): 997-1005.

Cohen, J. 1977. Statistical Power Analysis for the Behavioral Science. New York: Academic Press. Revised Edition.

Dupuy, B. (2000). Content-based instruction: Can it help ease the transition from beginning to advanced foreign language classes? *Foreign Language Annals*, 33(2): 205-223.

Echevarria, J., Short, D., and Powers, K. 2006. School reform and standards-based education: a model for English-language learners. Journal of Educational Research, 99(4): 195-210.

Echevarria, J., Short, D., and Powers, K. 2008. Making content comprehensible for non-native speakers of English: The SIOP model. International Journal of Learning, 14(11): 41-49.

Echevarria, J., Richards-Tutor, C., Chinn, V. P., and Ratleff, P. 2011. Did they get it? The role of fidelity in teaching English learners. Journal of Adolescent & Adult Literacy, 54(6): 425- 434.

Echevarria, J., Richards-Tutor, C., Canges, R. and 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.

Guarino, A.J., Echevarria, J., Short, D., Schick, J.E., Forbes, S. & Rueda, R. 2001. The Sheltered Instruction Observation: Reliability and validity assessment. Journal of Research Education, 11 (1):138-140.

Honigsfeld, A. and Cohan, A. 2008. The power of two: Lesson study and SIOP help teachers instruct ELLs. Journal of Staff Development 29(1): 24-28.

Jennions, M. and Moller, A. 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. Behavioral Ecology 14(3): 438-445.

Krashen, S. 1991. Sheltered subject matter teaching. Cross Currents 18: 183-188. Reprinted in J. Oller (Ed.) Methods That Work. Boston: Heinle and Heinle. pp. 143-148.

Krashen, S. 1994. The input hypothesis and its rivals. In N. Ellis (Ed.) Implicit and Explicit Learning of Languages. London: Academic Press. pp. 45-77.

Krashen, S. 2003. Explorations in Language Acquisition and Use: The Taipei Lectures. Portsmouth, NH: Heinemann.

Krashen, S. 2007. Case histories and the comprehension hypothesis. Selected Papers from the Sixteenth International Symposium on English Teaching, English Teachers' Association – Republic of China. Taipei: Crane Publishing Company. pp. 100-113.

McIntyre, E., Kyle, D., Chen, C., Muñoz, M., & Beldon, S. 2010. Teacher learning and ELL

reading achievement in sheltered instruction classrooms: Linking professional development to student development. Literacy Research & Instruction, 49(4), 334-351.

Morris, S. 2008. Estimating Effect Sizes from Pretest-Posttest-Control Group Designs. Organizational Research Methods. 11: 364-386.

Short, D., Echeverria, J., and Richards-Tutor, C. 2011. Research on academic literacy development in sheltered instruction classrooms. Language Teaching Research 15(3): 363-380.

Short, D., Fidelman, C., and Louguit, M. 2012. Developing academic language in English language learners through sheltered instruction. TESOL Quarterly 46(2): 334-361.

Truscott, J. 1996 The case against grammar correction in L2 writing classes. Language Learning 46(2): 327-369.

Truscott, J. 1999. What's wrong with oral grammar correction? The Canadian Modern Language Review 55(4): 437-56.

Welkowitz, J., Ewan, R. and Cohen, J. 1982. Introductory Statistics for the Behavioral Sciences. New York: Academic Press. Third edition.

Whittier, L. and Robinson, M. 2007. Teaching evolution to non-English proficient students by using Lego robotics. American Secondary Education 35(3): 19-28.