

**A Randomized Study of the Effects of Scaffolded Guided-Inquiry Instruction on  
Student Achievement in Science**

By

Rick Vanosdall  
Tennessee State University

Michael Klentschy  
El Centro (CA) Unified School District

Larry V. Hedges  
Northwestern University

Kathryn Sloane Weisbaum  
Tennessee State University

Paper presented at the Annual Meeting of the  
American Educational Research Association

April, 2007  
Chicago, Illinois

# **A Randomized Study of the Effects of Scaffolded Guided-Inquiry Instruction on Student Achievement in Science**

## **Introduction**

The use of inquiry-based teaching strategies has long been emphasized as the foundation of systemic reform initiatives in elementary science education. Reform efforts have focused on training teachers or providing materials, or both; on isolated schools or districts or states. Regardless of approach, the ultimate aim of these reforms is to improve student learning and achievement in science. However, the research on the effectiveness of these reform efforts—at least in terms of student achievement—has been spotty and inconclusive. In recent years, there has been a call for carefully designed experimental research to provide stronger tests of the causal effects of reform strategies on student achievement. In this paper, we report on a series of experimental and quasi-experimental studies designed to test the effects of a scaffolded guided-inquiry instructional system on student achievement. We compare these effects to those from other instructional systems, specifically kit-based instruction and textbook-based instruction.

## **Background**

While it is understood that educational reform efforts must take place within larger systemic and sociological contexts, the heart of inquiry-based instruction lies in the classroom. In reviewing the research literature on the effects of inquiry-based instruction on student achievement, we focus on three interrelated areas: 1) instructional materials; 2) instructional strategies; and 3) professional development to prepare teachers to use inquiry-based materials and/or strategies.

## **Instructional Materials**

Instructional materials designed for “hands-on” or “activity based” learning have been available since the first generation of kit-based curricula in the 1960’s and 1970’s. Subsequent meta-analyses of those early federally-supported programs indicated that some did improve student achievement in science. Bredderman (1983) synthesized research on the effectiveness of the three major activity-based science programs (Elementary Science Study, Science-A Process Approach, and Science Curriculum Improvement Study) that were in use by approximately one-fourth of the elementary teachers in the US during the 1976-77 school year (Weiss, 1978, cited in Bredderman, 1983). Compared to traditional textbook curricula, the overall mean effect size for activity-based curricula across all outcomes was 0.35, with mean effect sizes of 0.52 for science process tests and 0.16 for science content tests. Mean effects were noticeably stronger for disadvantaged students compared to other groups. And, mean effects were higher when the tests mirrored the activity-based approach (although other tests also resulted in positive, but lower, effects).

Using more refined statistical methods, Shymansky, Hedges, and Woodworth (1990) conducted a synthesis of the results of studies of the effects of new “inquiry oriented” science curricula used during the post-Sputnik era in K-12 classrooms. The inquiry oriented curricula that they examined were defined as those programs which

*emphasized the nature, structure, and processes of science; integrated laboratory activities into the core of instruction; and emphasized higher cognitive skills and an appreciation of science. Traditional curricula were defined as those programs which. . .emphasized knowledge of scientific facts, laws, theories and applications; and used laboratory activities as verification exercises or as lesson supplements.* (Shymansky, et.al., 1990, pg. 131).

Across all (K-12) science curricula, mean effect sizes favoring the new curricula were around 0.30 on the criteria of academic achievement and of process skills. These results suggest that, on the whole, students learned science content as well as process skills, contrary to critics' concerns that the new curricula emphasized process and attitude at the expense of content knowledge. However, estimates of effect sizes for achievement did vary substantially across elementary (K-6) versus secondary (7-12) grade levels, and across subject domains (e.g., chemistry had lower mean effects than biology or physics).

As in Bredderman's report, the authors noted serious deficiencies in the research designs and reporting methods in a large proportion of the individual studies reviewed. Specifically, the research designs of the evaluations studies that Shymansky, et al. examined varied considerably in quality, ranging from rather poorly controlled quasi-experiments to a handful of randomized experiments. Thus conclusions about these effect size estimates should be constrained by the fact that many of the studies on which they are based would not meet current standards for trustworthiness.

### **Instructional Strategies**

With the publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983) and subsequent "standards" documents (e.g., American Association for the Advancement of Science, 1990; National Research Council, 1996), reform efforts shifted to a greater focus on *instructional strategies* rather than the use of specific curricula. Inquiry-based, hands-on/minds-on, student-centered, materials-rich classroom instruction would enable students to construct understanding through experimentation, investigation, and interactions with teachers and peers (National Research Council, 1996). Proponents of standards-based education advocated "constructivist" instructional strategies and philosophies that transcended any given curricular package. The theoretical bases for such approaches were strong, but there were very few studies that examined empirically the impact of these strategies on student achievement.

In one attempt to examine the relationships between standards-based instructional strategies and student achievement, Von Secker and Lissitz (1999) analyzed data from the High School Effectiveness Study (HSES) that was part of the larger National Educational Longitudinal Study of 1988. Using hierarchical linear modeling techniques, they estimated the effects of inquiry-based instruction on (10<sup>th</sup> grade) student achievement in science. The strongest empirical support was observed for instruction that emphasized laboratory inquiry, but results were mixed for other instructional strategies. In a more detailed analysis, Von Secker (2002) calculated "effect sizes" (in the form of standardized regression coefficients) for instructional practices consistent with Standards recommendations. Controlling for student demographic status, she found that

*[o]n average, science achievement of a class increased by 0.58 standard deviation for every 1 standard deviation increase in the amount of emphasis that*

*their teachers placed on an inquiry approach that combined the following five practices, namely, a) eliciting student interest and engagement, b) using appropriate laboratory techniques, c) problem solving, d) conducting further study, and e) scientific writing. (p. 158)*

Von Secker also calculated estimates of the effects of each strategy individually. She found that: a) a one standard deviation increase in eliciting student engagement was associated with a 0.22 standard deviations increase in science achievement; b) a one standard deviation increase in use of appropriate laboratory techniques was associated with a 0.28 increase in science achievement; c) a one standard deviation increase in the use of problem solving as associated with a 0.33 standard deviation increase in science achievement; d) a one standard deviation increase in further study was associated with a 0.36 standard deviation increase in science achievement; and e) a one standard deviation increase in the use of scientific writing was associated with a 0.22 standard deviation increase in science achievement.

Note, however, that the data used in these studies were not based on observations of teacher behavior but on teacher surveys and self-reports of instructional practices. The literature on the validity of teacher self reports of instructional behavior suggests that observers' reports and teachers' reports of instructional behaviors do not always agree very well (see, e. g., Lee, et al, 2004; Schneider, et al, 2005).

Some researchers have found positive relationships between indicators of standards-based teaching practice and student achievement in elementary science (e.g., Kahle, et al. 2000), but others have not (e.g., Lawrenz, 2002; Lee et al 2004). And no studies that we are currently aware of, at least at the elementary school level, adequately control for volunteerism, selection bias, and other serious threats to the generalizability of results.

But if these standards-based instructional strategies *are* effective in helping students learn science, then a critical component of reform must be to ensure that teachers know how to use these strategies effectively. Many recent curricula are based on standards and designed to be inquiry-driven, such as those provided by the Full Option Science System (FOSS), Science and Technology for Children (STC), Science Education for Public Understanding Project (SEPUP), etc. However, even when inquiry-driven curriculum materials are available, inquiry-driven teaching differs from conventional teaching and presents substantial instructional challenges. Teachers need to learn how to use the materials and to manage the classroom experiments and student interactions in ways that may be novel to those teachers.

### **Teacher Professional Development**

Professional development programs aim to help teachers improve their knowledge of science content, improve their pedagogical knowledge of inquiry-based instruction in science, and give them confidence in using instructional methods that may be new for many teachers, particularly at elementary grade levels and in the subject of science (Lee, Hart, Cuevas, and Enders, 2004). Theoretically, teacher development helps teachers to improve their classroom instruction, which, in turn, promotes improved student achievement (Supovitz and Turner, 2000). In spite of the ubiquity of teacher development programs in science reform efforts, there is little research demonstrating a

convincing link between teacher development focusing on inquiry oriented science teaching and increased student achievement in science (Loucks-Horsley and Matsumoto, 1999; Porter, et al , 2000; Shymansky, et al 2004).

Federally funded initiatives to promote standards-based instruction in science placed a heavy emphasis on professional development. In Local Systemic Change (LSC) programs, approved lists of curricular materials (kit-based) were to be used in conjunction with 100 in the beginning and later 130 hours of professional development experience. A common evaluation framework (including required training for local evaluators) was developed in order to combine results across individual projects. Student achievement data, however, were not included in the evaluation framework and the data collection was not set up to include student data until the fourth cohort was funded. Results from the fourth cohort are not yet available.

In the “capstone” report summarizing the results across LSC projects, Banilower et al (2005) did find evidence that LSC projects influenced the content taught and the teaching strategies and materials used by participating teachers. However, almost none of the teachers were able to achieve the targeted number of professional development hours. And, when teachers did not know how to incorporate inquiry teaching strategies into their classroom practice, even the availability of inquiry-based instructional materials could not compensate totally for this lack of knowledge:

*According to evaluations, teachers in early stages of learning were more likely to use the materials mechanically, or to modify them in inappropriate ways. Some teachers ‘jumped the gun,’ skipping to activities on higher level concepts without adequate foundation for students. Still others omitted ‘rich’ activities, revised lessons, or added supplementary materials that shortcut the development of conceptual understanding as laid out in the curriculum. . . . Teachers’ lack of content knowledge also seemed to limit their use of materials in appropriate ways. As a result, while many teachers engaged students with the materials, the lessons had little student to student discussion and very limited questioning to promote concept development around the materials. (p. 55).*

When teachers lacked the confidence and experience in using investigative modeling, they tended to resort to more familiar “teacher-directed” instruction, overriding the intent of the materials. While some teachers can effectively use inquiry-based instruction regardless of the science curriculum they are using (textbook or otherwise), other teachers will use more “traditional” instruction regardless of the science curriculum they are using (kit based or otherwise). And, staff developers and researchers have certainly noted that teachers’ participation in professional development does not guarantee subsequent changes in teaching practice (e.g., Lee, Hart, Cuevas, and Enders, 2004; Schneider et al 2005).

The combined effect of curriculum materials and professional development has shown an impact on student achievement in some studies (Marx et al, 2004; Cuevas, et al 2005). It may be that a threshold of number of hours of participation must be reached (80-100) and time given for implementing change in the classroom (National Research Council, 2006; Shymansky, et al 2004).

But research has not always found that professional development has any benefit on student achievement over and above the effect that is achieved by curriculum materials alone. For example, Young and Lee (2005) compared the science achievement of 5<sup>th</sup> graders in districts that had a kit-based science curriculum supported by intensive professional development to 5<sup>th</sup> graders from other districts that did not use kit-based materials and did not have systematic science professional development for their teachers. Students in kit-based science classrooms scored significantly higher on a science achievement test than did students in non-kit-based classrooms.<sup>1</sup> But there was no difference in the posttest performance of students in kit-based classrooms whose teachers who had a high number of professional development hours compared to those with a low number of professional development hours. Further, there were no systematic differences in the instructional strategies of teachers with high vs. low number of professional development hours. It is perhaps not the participation per se, but rather whether standards based teaching practices are used in conjunction with appropriate curricular materials that make a difference in student achievement (Kahle, et al 2000).

### **The Development of Scaffolded Guided Inquiry**

Instructional practices can be conceptualized as falling along a continuum from complete open inquiry (defined as “investigations [or] free-ranging explorations of unexplained phenomenon” (National Research Council, 2001)) to teacher demonstration. Degrees of inquiry, either “guided” or “directed” fall along this continuum, nearer or farther from open inquiry, respectively.

Most investigative science instruction in today’s standards-based classrooms is guided inquiry. Guided inquiry provides teachers with the opportunity to carefully plan classroom investigations that will both expose students to the required science content (standards) and plan the implementation of lessons that will integrate student science notebooks and classroom discussion as a means for students to develop a deep understanding of the required content. Ideally, there is alignment of the science content standards that should be taught (intended curriculum) with what is actually taught (implemented curriculum) with what is actually learned by students (achieved curriculum) (Marzano, 2003, 2001).

Research on how students learn science recognizes that the development of deep conceptual understanding in science takes time and is enhanced through providing supports, scaffolds or prompts to guide students to enhance their scientific reasoning

---

<sup>1</sup> There were significant pretest differences between the two groups that were not controlled for in the analyses of posttest scores, which seriously compromises the interpretation of the student achievement data in this study. Moreover, the test used in this study was that developed by Horizon Research, Inc., from National Assessment of Educational Progress (NAEP) and Third International Mathematics and Science Survey (TIMSS) items, and used for the core evaluation of NSF-funded LSCs. This test was designed to measure the objectives of the kit, but may not have been well aligned with the instruction students received in the comparison group

ability (National Research Council, 2005). Thus, the classroom teacher needs to guide the inquiry process in order to develop both deep conceptual understanding and scientific reasoning ability to formulate explanations based on evidence. But guided inquiry is a complex process because students and teachers may often lack content knowledge, inquiry experience, and resources and are unable to make meaningful inferences from data without adequate support.

Recognizing these limitations, Klentschy and others (Klentschy, 2005, 2004; Amaral, Garrison, and Klentschy, 2002) revisited the notion of alignment with a level of guided inquiry that added scaffolds or supports for students within this level of inquiry. In their approach, students are guided and supported through the process of constructing their understanding of scientific concepts and the process of scientific inquiry as they work through the lessons, record predictions, observations, and reflections in their journals, and learn to articulate claims and evidence for their conclusions.

“Scaffolding” occurs for teacher learning, as well. The teachers’ guides are modified in several important ways, to model for teachers the essential elements of effective standards-based instruction. First, the lessons in the unit are linked directly to specific standards in the state curriculum and assessment guides. Teachers know what standards are being addressed in each unit and lesson. Second, critical or “benchmark” lessons are identified so the teachers know which lessons are critical in the development of student understanding (and therefore don’t ‘jump the gun’ as described by Banilower et al, 2005). Third, questioning, experimentation, and reflection are all modeled in order to support the teacher through classroom activities and interactions. Finally, the use of student notebooks is emphasized as a way for the teacher to assess student’s understanding and to provide the feedback that is necessary for student learning.

Scaffolded guided-inquiry lessons were developed for the FOSS unit on Mixtures and Solutions, a unit which reflects the content standards for 5<sup>th</sup> grade science learning in California. In a series of experimental studies, the scaffolded guided-inquiry lessons were compared to existing FOSS units (“kit-based”) and to textbook based instruction in a series of experimental studies designed to test the relative effectiveness of each method on student achievement in science.

### **Context of the Studies**

The research reported here is part of a larger multi-year, multi-site project on “Academic Achievement and Teacher Development in Science (AATDS),” supported by the Interagency Educational Research Initiative (IERI). One purpose of the project is to determine whether well conceived, systematic, and extensive teacher development, in conjunction with standards-based science curricula, can have a causal effect on student achievement. The studies described in this paper were designed to explore the effects of scaffolded lessons for guided science inquiry instruction for teachers with and without previous experience in kit-based instruction.

The site for these studies is the Imperial Valley in California. Imperial County is in the southeast corner of California along the United States border with Mexico. The county is one of the largest and most sparsely populated counties in California, and many residents live in extreme poverty with a mean per capita income the lowest of all California counties. Of the 22,500 K-6 students in the Imperial Valley, 81% are Hispanic, 5% African American, 11% Caucasian, 1% Asian and 1% Native American.

More than 50% of the students are Limited English Proficient, with 10% children of migrant workers. Nearly all of the county's schools qualify for Title I. County-wide, more than 67% of all students are eligible for free and reduced lunches. There are 14 elementary school districts in Imperial County. Five of these districts participated in the studies described below. The five participating districts reflect the county averages in the distributions of Hispanic students (71%-98%), eligibility for free or reduced lunches (36%-86%), and English Language Learners (23%-71%)<sup>2</sup>.

The data reported here were collected during the 2004-05 academic year. All participating classes are 5<sup>th</sup> grade science classes with average class sizes (district level) ranging from 20.6 to 26.5.

### ***Study 1: Randomized Study of the Effects of Scaffolded Guided Inquiry Compared to Textbook-Based Instruction***

#### **Rationale and Research Questions**

This first study was designed to determine if use of the scaffolded guided inquiry approach results in stronger student performance in science, compared to the use of traditional textbook-based curricula. Two of the districts in Imperial County had not previously participated in systematic reform efforts in science education (such as the Local Systemic Initiative or VIPS, described in Study 2). Fifth grade science teachers were using the textbook series that had been adopted by the districts (Harcourt Brace in one district and McGraw Hill in the other). This provided the ideal setting to examine the effects of this new inquiry-based approach (scaffolded guided-inquiry) to science instruction with teachers who had no previous experience with kit-based curricula.

#### **Design**

All four elementary schools across the two districts agreed to participate in the study. Fifth grade teachers within each school were asked to participate in the experiment. Those who agreed to participate were randomly assigned to be in either the treatment or comparison groups for this study. Those teachers who were assigned to the comparison group were offered the option to be trained in the new techniques after the conclusion of the study, if they wished.

#### **Sample**

The sample consists of 20 fifth grade teachers and 563 students across four schools. Ten teachers were randomly assigned to the treatment group and ten were assigned to the comparison group.

#### **The Treatment Condition: Scaffolded Guided Inquiry-Based Instruction (SGI)**

Teachers in the treatment group received the FOSS kits on Mixtures and Solutions. This curriculum unit was selected because the California state content standards specifically assess the content covered in this unit. In place of the standard Teachers' Guide accompanying the kit materials, the teachers received the Scaffolded Lessons Guide. They participated in 6 hours of professional development focusing on the use of the kit materials and the scaffolded lessons. Training occurred during February,

---

2005 and the teachers taught the unit (6-8 weeks) during the spring semester of 2005. During this instruction period, treatment teachers received in-classroom support and coaching pertaining to the unit and the use of scaffolded lessons from the project Science Resource Teachers. All teachers received the same regular classroom support as being a part of VIPS project.

### **The Control Condition: Textbook-Based Instruction**

Teachers in the comparison group used their standard textbook curriculum and whatever “off the shelf” materials they typically used in their science classrooms. Their professional development in this curriculum consisted of the training offered by the respective publishing companies when the textbook series were adopted by the district. Their classrooms were also visited by Science Resource Teachers and they had access to that the same regular classroom support as being a part of VIPS project.

### **Instruments**

A wide variety of school, teacher background, classroom practice, and student achievement data were collected in the course of the study, but this report focuses primarily on the relation of treatment group assignment (scaffolded guided inquiry-based instruction versus text-based instruction) to student achievement. We used two measures of student achievement, one intended to be well aligned to the specific content of the instructional unit and one intended to be more distal from the unit but more relevant for accountability purposes.

The achievement measure designed to be well aligned with instruction was the FOSS unit test. Students in the treatment group were administered one form of the FOSS Unit Test as a pre-test, before instruction in the unit began, and another form as a post-test at the end of unit instruction. Students in the comparison group were administered the FOSS Unit Test (California Edition) as a pretest before instruction started for their science unit and as a posttest when instruction ended for their science unit.

The achievement measure designed to be more distal was the California Standards Test (CST), which is a 70 item standardized test used for state accountability purposes. It was administered in the Spring of the students’ 5<sup>th</sup> grade year, according to state schedules and guidelines. Only the physical sciences subtest scores were used in this study. Although the state does not set a formal proficiency standard for any subscale, a score of 6 on each subscale (out of 11 on the physical science subscale) would lead to an overall score level in the proficient range. Thus the score 6 can be used as a rough guide to the state performance level identified as proficient.

### **Analyses**

The data analysis involved a two level hierarchical linear model (HLM) with students nested within classrooms (see Raudenbush and Bryk, 2002). The specific model for achievement test score  $Y_{ij}$  of the  $i^{\text{th}}$  student in the  $j^{\text{th}}$  classroom (the level one model) was

$$Y_{ij} = \beta_{0j} + \beta_{1j}PRETEST_{ij} + \varepsilon_{ij},$$

where  $PRETEST_{ij}$  is a pretest score,  $\beta_{0j}$  is the (covariate adjusted) classroom mean,  $\beta_{1j}$  is the effect of the covariate in the  $j^{\text{th}}$  classroom, and  $\varepsilon_{ij}$  is a student-specific residual.

The specific model for variation of coefficients between classes (the level 2 model) was

$$\beta_{0j} = \gamma_{00} + \gamma_{01}TREATMENT_j + \eta_{0j},$$

and

$$\beta_{1j} = \gamma_{10},$$

where  $\gamma_{00}$  is (the average covariate adjusted mean of the control group),  $TREATMENT_j$  is a dummy variable for treatment (SGI),  $\gamma_{01}$  is the treatment effect,  $\gamma_{10}$  is the effect of pretest on posttest (which, is fixed across classrooms because it is used as a covariate), and  $\eta_{0j}$  is classroom-specific random effect. Thus the object of the analysis is to estimate the three fixed effects ( $\gamma_{00}$ ,  $\gamma_{01}$ , and  $\gamma_{10}$ ) as well as the variance of the covariate adjusted classroom means (the variance of the  $\eta_{0j}$ 's).

Because teachers (and their classrooms) were randomly assigned to treatments, no pretest differences were anticipated. However, to check that there were no unanticipated pretest differences, we carried out a hierarchical linear model analysis using essentially the same model given above, except that pretest was the dependent variable (that is  $Y_{ij}$  was the pretest) and pretest was not included as a covariate in the level one model.

## Results

The means and standard deviations for the treatment (SGI) and control (textbook-based instruction) groups on the three measures of student achievement are presented in Table 1. The mean pretest scores for the two groups were nearly identical (10.470 for the treatment group and 10.440 for the comparison group). This corresponds to a pretest mean difference of less than 0.01 standard deviations. Our HLM analysis of pretest mean differences confirmed that the pretest means were not statistically significantly different (see Table 2).

Students in classrooms using traditional textbook-based instruction had an average gain of less than 1 point from the pretest to the posttest. This corresponds to a gain of about 0.2 pretest standard deviations. In contrast, students in the SGI classrooms had an average gain of over 6 points on the FOSS unit posttest, compared to their pretest performance, corresponding to a gain of about 1.5 pretest standard deviations. Comparing the SGI and textbook-based instruction groups on the FOSS unit posttest, we see that the mean posttest score of the SGI group is over 5 points higher than that of the textbook-based instruction groups, which corresponds to a difference of about 1.4 standard deviations in favor of the SGI group. The SGI group also had a mean on the California Standards Test that was about 2.5 points higher than that of the textbook-based instruction group, which also corresponds to about 1.4 standard deviations in favor of the SGI group.

The results of the HLM analyses show that the treatment effect (SGI versus textbook-based instruction) on both the FOSS unit test (Table 3) and the physical science subtests of the California Standards Test (Table 4) are positive and statistically significant. The estimated treatment effects shown in these tables are adjusted for students' scores on the FOSS unit pretest. These pretest-adjusted treatment effects correspond to effect sizes of 1.09 standard deviations on the FOSS unit test and 1.39 standard deviations on the California Standards Test.

## **Discussion of Study 1**

Study 1 demonstrated that the use of scaffolded guided inquiry based instruction in combination with kit-based curriculum led to higher science achievement than did textbook-based instruction. These differences were evident both in tests that were aligned with the content of the instruction (the FOSS unit test) and with the state assessment used for accountability purposes (the CST). Moreover, the benefits of scaffolded guided inquiry based instruction were substantial. The effect sizes obtained (in excess of one standard deviation) are large by any standard of comparison. For example, in Cohen's (1977) classification of effect sizes (based on effects typically found in the social sciences) as small, medium, or large, an effect size greater than 0.8 is classified as large. Similarly, the average amount of gain between the beginning of grade 5 and the beginning of grade 6 (one year's growth) on most standardized tests is less than one standard deviation. Thus we could interpret the treatment effect as being as large as one year's worth of growth (which includes the effect of development and maturation as well as everything that is learned both in and out of school).

In another vein, the average posttest score on the California Standards Test is just over 6 in the group that received scaffolded guided-inquiry instruction but is less than 4 in the text-based instruction group. This has the important practical interpretation that the average student receiving scaffolded guided-inquiry instruction would (roughly) fall into the range declared proficient on the California Standards Test, while the average student in the text-based instruction group would fall far short of that.

One surprising aspect of these results is that the treatment effect is somewhat larger on the California Standards Test than on the FOSS unit test. We might have expected just the opposite, namely, larger treatment effects on a test that is linked specifically to the experimental curriculum (i.e., the FOSS kit). The reason for the (slightly) larger treatment effects on the California Standards Test may lie in the design of the SGI treatment. The scaffolding provided for each lesson was specifically targeted at one of the California state standards. Aspects of the FOSS unit that were not emphasized in California state standards were not emphasized in the scaffolded lessons, and likely received little or no emphasis in instruction in treatment classrooms. The California state assessment, in turn, was designed to measure achievement of those state standards. Consequently the scaffolded lessons may have led to instruction that was better aligned with the California's State assessment than with the FOSS unit test.

## **Study 2: Randomized Study of Effects of Scaffolded Guided Inquiry Compared to Kit-Based Instruction**

### **Rationale and Research Questions**

The results of Study 1 demonstrate that scaffolded guided inquiry-based instruction, in conjunction with kit-based materials lead to higher student achievement in science than did text-based instruction, but leave open the question of whether it is the kit-based curriculum or the scaffolded lessons that are producing such a dramatic improvement in achievement. To determine which component of the treatment in Study 1 was responsible for the effects, it was necessary to carry out another study that separated these components. Because previous research (e.g., Bredderman, 1983;

Shymansky, et al., 1990) suggested much smaller effects for kit-based materials alone, we hypothesized that the scaffolded guided instruction component was responsible for a substantial portion of the large effects we observed in Study 1. Consequently, we conducted a second experiment designed to determine whether scaffolded guided instruction using kit-based materials led to greater student achievement in science than did the use of kit-based materials alone.

Study 2 involved a group of teachers from three school districts who had experience in the use of kit-based materials in inquiry-based science education. The three districts involved in this study had been part of the Valle Imperial Project in Science (VIPS), which arose from a NSF funded Local Systemic Initiative. VIPS began in 1998 as a collaborative partnership between Imperial County school districts and San Diego State University, Imperial Valley Campus. One of the districts in this study also served as the pilot site three years prior. Participating VIPS teachers had received high quality inquiry-based curricular materials in the form of kits or modules drawn from sources such as 1) Science and Technology for Children (STC) developed by the National Science Resource Center (NSRC) at the Smithsonian Institute supported by the National Academy of Science; 2) Full Option Science System (FOSS) developed at the Lawrence Hall of Science, University of California, Berkeley; and 3) Insights created by the Education Development Center in Newton, Massachusetts. In addition to materials, teachers participating in VIPS had been provided with professional development designed to deepen their own content understanding and address pedagogical issues.

### **Design**

Fifth grade teachers who were previously VIPS participants were matched across schools on relevant background variables (years experience teaching in the system; number of hours of professional development; and experience with kit-based curriculum, in that order). One of each of the matched pairs was then randomly assigned to the control group (continue current strategies of kit-based instruction), and the other to the treatment group (receive the scaffolded lesson materials to use in conjunction with kits). As in Study 1, comparison teachers were given the option of receiving the scaffolded lessons and training after the conclusion of the study.

### **Sample**

Fifth grade teachers from 11 schools participated. The 24 teachers were matched into 12 pairs and then one of each pair was randomly assigned to treatment ( $n = 12$ ) or comparison groups ( $n = 12$ ). Class sizes ranged from a high of 37 to a low of 23 students per class. A total of 762 students participated.

### **Treatment**

Teachers in the treatment group received the FOSS kit on Mixtures and Solutions, the Scaffolded Lesson Guide, and training in the use of this approach. Teachers in the control group received the FOSS kit on Mixtures and Solutions, but not the Scaffolded Lesson Guide or any additional training. Both treatment and control group teachers received the normal class support available to all VIPS participants. Training occurred in the Fall of 2004 and the units were taught in the second and third trimester of the 2004-2005 school year as part of a regular rotation.

## **Instruments**

As in Study 1, a wide variety of school, teacher background, classroom practice, and student achievement data were collected in the course of the study. In this paper, we report only results based on the data collected on two achievement tests. As in study 1, one of the achievement tests was intended to be closely aligned with the curriculum (the FOSS unit test) and the other was intended to be more distal but more relevant for accountability purposes (the California Standards Test). Students in the treatment group were administered one form of the FOSS Unit Test as a pre-test, before instruction in the unit began, and another form as a post-test at the end of unit instruction. Students in the comparison group were administered the FOSS Unit Test (California Edition) as a pretest before instruction started for their science unit and as a posttest when instruction ended for their science unit. The California Standards Test was administered in the Spring of the students' 5<sup>th</sup> grade year, according to state schedules and guidelines. Only the physical sciences subtest scores were used in this study.

## **Analyses**

The data analysis utilized the same two level hierarchical linear model described in connection with Study 1, except that here the control condition was kit-based instruction.

## **Results**

The means and standard deviations for the treatment (SGI) and control (kit-based instruction) groups on the three measures of student achievement are presented in Table 5. The mean pretest scores for the two groups were nearly identical (10.79 for the treatment group and 11.06 for the control group). This corresponds to a pretest mean difference of about 0.08 standard deviations favoring the control group. However, our HLM analysis of pretest mean differences confirmed that the pretest means were not statistically significantly different (see Table 6).

Students in classrooms using kit-based instruction had an average gain of about 1.6 points from the pretest to the posttest on the FOSS unit test. This corresponds to a gain of about 0.5 pretest standard deviations. In contrast, students in the SGI classrooms had an average gain of almost 6 points on the posttest, compared to their pretest performance, corresponding to a gain of about 1.75 pretest standard deviations. Comparing the SGI and kit-based instruction groups at posttest, we see that the mean FOSS unit test score of the SGI group is about 4 points higher than that of the kit-based instruction groups, which corresponds to a difference of about 1.0 standard deviations in favor of the SGI group. The SGI group also had a mean on the California Standards Test that was about 2.2 points higher than that of the kit-based instruction group, which also corresponds to about 1.3 standard deviations in favor of the SGI group.

The results of the HLM analyses show that the treatment effect (SGI versus kit-based instruction) on both the FOSS unit test (Table 7) and the physical science subtests of the California Standards Test (Table 8) are positive and statistically significant. These estimated treatment effects are adjusted for the students' scores on the FOSS unit pretest. These pretest-adjusted treatment effects correspond to effect sizes of 1.04 standard deviations on the FOSS unit test and 1.14 standard deviations on the physical sciences

subtest of the California Standards Test. Once again, the scaffolded guided inquiry group outperformed the comparison group. It appears that the scaffolded lessons add an important and significant “additive effect” to the kit-based instruction.

### **Discussion of Study 2**

Study 2 demonstrated that the use of scaffolded guided inquiry based instruction in combination with kit-based curriculum led to higher science achievement than did kit-based instruction alone. These differences were evident in both tests that were aligned with the content of the instruction (the FOSS unit test) and with the state assessment used for accountability purposes (the CST). Moreover, the benefits of scaffolded guided inquiry based instruction as compared to kit-based instruction alone were substantial. As in the case of Study 1, the effect sizes obtained (in excess of one standard deviation) are large by any standard of comparison, such as Cohen’s (1977) classification of effect sizes. The average amount of gain between the beginning of grade 5 and the beginning of grade 6 (one year’s growth) on most standardized tests is less than one standard deviation. Thus we could interpret the effect of supplementing kit-based instruction with scaffolded guided instruction as being as large as one year’s worth of growth (which must be interpreted as including the effects of maturation and whatever is learned in and out of school).

As in the case of Study 1, the scores obtained on the California Standards test have important practical implications in terms of the state accountability system. The average posttest score on the California Standards Test is just over 6 in the group that received scaffolded guided-inquiry instruction but is less than 4 in the kit-based instruction group. Therefore the average student receiving scaffolded guided-inquiry instruction would (roughly) fall into the range declared proficient on the California Standards Test, while the average student in the kit-based instruction group would fall short of that.

## **Study 3: Combined Study to Compare Kit-Based Instruction with Textbook-Based Instruction**

### **Rationale and Research Question**

Study 1 demonstrated that scaffolded guided inquiry used in conjunction with kit-based materials lead to higher student achievement than traditional textbook-based instruction. Study 2 demonstrated that scaffolded guided inquiry in conjunction with kit-based materials led to higher student achievement than instruction using kit-based materials alone. This raises the question of whether instruction using kit-based materials alone would lead to higher achievement than textbook-based instruction.

Perhaps the strongest approach to answering this question would be to conduct a randomized experiment assigning some teachers (and their classrooms) to provide instruction using kit-based materials and others to provide textbook-based instruction. Another (albeit weaker) approach to answering this question is to use the data from Studies 1 and 2 to construct a quasi-experimental comparison of the effects of kit-based instruction versus textbook-based instruction.

The treatment groups in Studies 1 and 2 were both assigned to use the scaffolded guided inquiry approach in conjunction with kits. The control group in Study 1 used traditional textbook-based instruction, and the control group in Study 2 used kit-based materials only. In Study 3 we combined the samples (specifically, the control groups) from the two studies in order to construct a quasi-experiment on kit-based instruction versus textbook-based instruction.

It is important to recognize that while Study 1 and Study 2 are randomized experiments and therefore provide estimates of treatment effects that are relatively free from bias, the results of this quasi-experiment are necessarily more equivocal because there was no random assignment of teachers to either the kit-based or the textbook-based instruction conditions. Therefore there is no assurance (based on randomization) of equivalence of the teachers in these two conditions (or their students). The assurance that the treatment and comparison groups were essentially identical except for the treatments being compared must come from other sources such as pretest data and other baseline covariates and the fact that the teachers in both studies came from adjacent districts in a rather homogeneous rural area.

### **Design**

The control groups in Studies 1 and 2 were kept intact, with 12 teachers and 370 students in the kit-based instruction group and 10 teachers and 287 students in the textbook-based instruction group. The treatment groups from Study 1 and Study 2 were not used in this quasi-experiment.

### **Analyses**

The data analysis involved a two-level hierarchical linear model (HLM) with students nested within classrooms that was similar to that used in Studies 1 and 2. The major difference is that, in this analysis, the treatment is kit-based instruction (without SGI) and the comparison condition is text-based instruction. Thus the dummy variable for treatment ( $TREATMENT_i$ ) is a dummy variable for kit-based instruction.

### **Results**

The means and standard deviations for the treatment (kit-based instruction) and comparison (text-based instruction) groups on the three measures of student achievement are presented in Table 9. The mean pretest scores for the two groups were nearly identical (11.06 for the group using kit-based materials, and 10.44 for the text-based instruction group). These means imply a pretest mean difference of about 0.6 or about 0.125 standard deviations. Our HLM analysis of pretest mean differences confirmed that the pretest means were not statistically significantly different (see Table 10).

Students in the group of classrooms using kit-based materials without SGI had an average gain of about 1.6 points from the pretest to the posttest on the FOSS unit test. This corresponds to a gain of just under 0.5 pretest standard deviations. The text-based instruction group has an average gain of 0.9 points from the pretest to the posttest on the FOSS unit test, which corresponds to just under 0.2 pretest standard deviations. Comparing the kit-based and text-based instruction groups at posttest, we see that the mean FOSS unit test score of the kit-based group is about 1.3 points higher than that of the text-based instruction group, which corresponds to a difference of about 0.3 standard

deviations in favor of the kit-based group. The difference on the physical science subscale of the California Standards Test is about 0.5 points, corresponding to an effect size of over 0.25 standard deviations.

The results of the HLM analyses show that the treatment effect (kit-based instruction versus text-based instruction) on both the FOSS unit test (Table 11) and the physical science subtests of the California Standards Test (Table 12) are positive and statistically significant. The estimated treatment effects are adjusted for students' scores on the FOSS unit pretest. These pretest-adjusted treatment effects correspond to effect sizes of 0.41 standard deviations on the FOSS unit test and 0.32 standard deviations on the physical sciences subtest of the California Standards Test.

### **Discussion of Study 3**

This quasi-experimental comparison suggests that even without the addition of scaffolded guided-inquiry, teachers with experience using kit-based materials may be able to achieve greater student achievement gains than those using text-based instruction. The effect for kit based materials on the FOSS unit test was positive and statistically significant. The effect of kit-based materials on the CST was positive and nearly statistically significant ( $p=.055$ ). The magnitude of the benefits we observed, about 0.4 standard deviation units, is consistent with the effects for kit-based instruction versus text-based instruction found in the meta-analysis by Shymansky, et al. (1990). Therefore our results appear to be consistent with those of earlier research on kit-based materials.

Note that the average posttest score on the California Standards Test is less than 4 in the kit-based instruction group. While this is larger than the average score in the text-based instruction group, it is nonetheless considerably less than would be required to yield an overall score of proficient on the state accountability standards.

## **Study 4: Combined Study to Compare Effectiveness of Experienced and Inexperienced Teachers Using SGI**

### **Rational and Research Questions**

Study 1 demonstrated that teachers with no prior experience using kit-based materials can produce greater student achievement when they use scaffolded guided-inquiry instruction in conjunction with kit-based materials, rather than using traditional text-based instruction. Study 2 demonstrated that teachers who had prior experience with kit-based materials can produce greater student achievement when they use scaffolded guided-inquiry instruction in conjunction with kit-based materials, rather than using kit-based materials alone. This raises the question whether prior experience with kit-based materials improves the effectiveness of teachers using scaffolded guided-inquiry instruction.

As in the case of Study 3, the strongest design to answer this question would be to carry out a randomized experiment that assigned some teachers to first obtain experience in kit-based instruction, others not to do so, and then compared the student achievement of the students of these two groups of teachers at a later point when they were both using scaffolded guided-inquiry instruction. Such an experiment would be expensive, time consuming, and difficult to implement (since it depends on teachers maintaining an assignment condition for years while obtaining experience using kit-based materials, or not doing so). A more feasible approach is to carry out a quasi-experiment by combining data on the treatment groups of Study 1 (teachers with no prior experience using kit-based materials) and Study 2 (teachers with prior experience using kit-based materials). In Study 4 we combined the samples (specifically, the treatment groups) from the two studies in order to construct a quasi-experiment to evaluate the impact of teacher experience with kit-based materials on their effectiveness in using scaffolded guided-inquiry.

As in the case of Study 3, Study 4 is a quasi-experiment that does not have the same logical status as the randomized experiments which provide the data used for the comparisons. There was no random assignment of teachers to either prior experience or no prior experience with instruction using kit-based materials. Therefore there is no assurance (based on randomization) of equivalence of the teachers in these two conditions (or their students). The persuasiveness of the conclusions of this study therefore depends on the persuasiveness that the two groups of teachers and students were equivalent except for the level of teacher experience with kit-based materials.

### **Design**

The treatment groups in Studies 1 and 2 were kept intact, with 12 teachers and 392 students in the “prior experience with kit-based materials” group and 10 teachers and 276 students in the “no prior experience with kit-based materials” group. The control groups from Study 1 and Study 2 were not used in this quasi-experiment.

### **Analyses**

The data analysis involved a two level hierarchical linear model (HLM) with students nested within classrooms that was similar to that used in Studies 1, 2 and 3. The major difference is that, in this analysis, the treatment is prior experience with kit-based instruction and the comparison condition is no prior experience with kit-based materials.

Thus the dummy variable for treatment ( $TREATMENT_i$ ) is a dummy variable for prior experience with kit-based instruction.

## Results

The means and standard deviations for the treatment (prior experience with kit-based instruction) and comparison (no prior experience) groups on the three measures of student achievement are presented in Table 13. The mean pretest scores for the two groups were nearly identical (10.79 for the group having prior experience with kit-based materials, and 10.47 for the group having no prior experience). These means imply a pretest mean difference of about 0.3 points or less than 0.1 standard deviations. Our HLM analysis of pretest mean differences confirmed that the pretest means were not statistically significantly different (see Table 14).

Students in the group of classrooms whose teachers had prior experience with kit-based materials had an average gain of almost 6 points from the pretest to the posttest on the FOSS unit test. This corresponds to a gain of about 1.75 pretest standard deviations. The students whose teachers had *no* prior experience with kit-based materials had an average gain of just over 6 points from the pretest to the posttest on the FOSS unit test, which corresponds to 1.5 pretest standard deviations. Comparing the groups of students whose teachers had prior experience with kit-based materials with the group who had *not* had prior experience at posttest, we see that the mean FOSS unit test score of the experienced group is about almost identical to that of the inexperienced group—the difference is less than 0.05 standard deviations. Also, there is essentially no difference on the physical science subscale of the California Standards Test: the difference corresponding to an effect size of only 0.01 standard deviations.

The results of the HLM analyses confirm that the treatment effects (of prior experience with kit-based instruction versus no such experience) are very small and are statistically insignificant on both the FOSS unit test (Table 15) and the physical science subtests of the California Standards Test (Table 16). Similarly, the effect sizes are very small, only about 0.13 standard deviations for both the CST and the FOSS posttest.

## Discussion of Study 4

This quasi-experiment demonstrates that prior experience with kit-based materials is not necessary to be effective in the use of scaffolded guided-inquiry instruction. This is an important finding in the context of science education reform policy. Many such policies advocate inquiry-based science instruction, but recognize that inquiry-based instruction requires methods that are somewhat different than other forms of instruction. They hypothesize that, for this reform to be effective, it is necessary to build the capacity for effective inquiry-based teaching through teacher development and experience with inquiry-based teaching. In this and other capacity-building reforms, one often expects that there will be little or improvement in academic achievement until teachers gain experience with the method. The results of this study suggest that scaffolded guided-inquiry may be an effective way to implement inquiry-based science teaching rapidly. To put it another way, scaffolded guided-inquiry may permit teachers with relatively little experience to implement inquiry-based science instruction as well as teachers with more experience in inquiry-based instruction using kit-based materials.

There seems to be little reason to believe that there are profound differences between the groups being compared except for their prior experience with kit-based materials in inquiry-based science instruction. The current literature on the effectiveness of quasi-experimental designs identifies two considerations that are important in ensuring that they are relatively free from bias (see, e.g., Heinsman & Shadish, 1996). The first consideration is the degree to which the baseline covariates match in the groups being compared (matching on observable individual characteristics). Our analyses show that the groups are very well matched on pretest scores. The second consideration is the degree to which the comparison groups are formed from local populations (which leads to likely matching on unobservable characteristics). The two studies whose data were combined in Studies 3 and 4 were conducted in adjacent school districts within the same rather isolated county. The student and teacher characteristics were very similar and the reason that some teachers had prior experience with inquiry oriented instruction using kit-based materials was that one district had participated in the LSC project and the other had not, thus there was essentially no selection to do so (or not) on the part of individual teachers.

### **Overall Conclusions**

Two randomized experiments demonstrated that scaffolded guided-inquiry used in conjunction with kit-based materials dramatically improved fifth grade science achievement compared to either text-based instruction or instruction using kit-based materials alone. Quasi-experimental evidence derived from these experiments suggests that while kit-based instruction in the hands of experienced teachers may be more effective than text-based instruction, the effects appear to be much smaller than those of scaffolded guided instruction. Additional quasi-experimental evidence derived from the experiments suggests that teachers with no prior experience with instruction using kit-based materials may be just as effective in using scaffolded guided-inquiry instruction (in conjunction with kit-based materials) as are teachers with prior experience using kit-based materials. This suggests that the use of scaffolding to guide inquiry-based instruction may make it possible to implement inquiry oriented reforms more rapidly and effectively.

The research findings presented in this paper are based on the instruction corresponding to a single instructional unit (FOSS Mixtures and Solutions) and a single grade level (grade 5). It is not obvious that these results would generalize to different units or different grade levels. Similarly, this research was conducted in a single, geographically isolated rural area having an economically disadvantaged student population with a large proportion of English language learners. It is possible that these results would not generalize to settings with more advantaged student populations. However we regard these findings as promising and are currently conducting additional studies to determine if these findings generalize to other grade levels and to other geographical settings with student bodies having different socioeconomic compositions.

## References

- Amaral, O., Garrison, L., & Klentschy, M. (Summer, 2002). Helping English learners increase achievement through inquiry-based science instruction. *Bilingual research journal*, 26(2), 213-239.
- American Association for the Advancement of Science (1990). *Science for all Americans*. NY: Oxford University Press.
- Banilower, E. R., Boyd, S. E., Pasley, J. D., & Weiss, I. R. (November 2005). *Lessons from a decade of mathematics and science reform: A capstone report for the local systemic change teacher enhancement initiative* (Prepared for the National Science Foundation No. Prepublication Copy, updated February 2006). Chapel Hill, N.C.: Horizon Research, Inc.
- Bredderman, T. (Winter, 1983). Effects of activity-based elementary science on student outcomes: A quantitative synthesis. *Review of Educational Research*, 53(4), 499-518.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. NY: Academic Press.
- Committee on Science and Mathematics Teacher Preparation, National Research Council. (2000). *Educating teachers of science, mathematics, and technology: New practices for the new millennium*. Washington, DC: National Academies Press. from <http://www.nap.edu/catalog/9832.html>
- Cuevas, P., Lee, O., Hart, J., & Deaktor, R. (2005). Improving science inquiry with elementary students of diverse backgrounds. *Journal of Research in Science Teaching*, 42(3), 337-357.
- Heinsman, D.T. & Shadish, W.R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychological Methods*, 1, 154-169.
- Kahle, J. B., Meece, J., & Scantlebury, K. (2000). Urban african-american middle school science students: Does standards-based teaching make a difference? *Journal of Research in Science Teaching*, 37(9), 1019-1041.
- Klentschy, M. (November/December, 2005). Science notebook essentials. *Science and Children*, 43(3), 24-27.

- Klentschy, M. & Molina-De La Torre, E. (2004). Students' science notebooks and the inquiry process. In W. Saul (Ed.), *Crossing borders in literacy and science instruction: Perspectives on theory and practice*. Newark, DE: International Reading Association Press.
- Lawrenz, F., & Huffman, D. (2002). *Science education reform: The impact of teacher enhancement and curriculum implementation on student performance, 1995-2001* (Report to the National Science Foundation(ERIC Document Service No. ED467639) Retrieved February 28, 2005, from ERIC database.
- Lee, O. (2004). Teacher change in beliefs and practices in science and literacy instruction with english language learners. *Journal of Research in Science Teaching*, 41(1), 65-93.
- Lee, O., Hart, J., Cuevas, P., & Enders, C. (2004). Professional development in inquiry-based science for elementary teachers of diverse student groups. *Journal of Research in Science Teaching*, 41(10), 1021-1043.
- Loucks-Horsley, S. & Matsumoto, C. (May, 1999). Research on professional development for teachers of mathematics and science: The state of the scene. *School Science and Mathematics*.
- Luykx, A., & Lee, O. (2006). Measuring instructional congruence in elementary science classrooms: Pedagogical and methodological components of a theoretical framework. *Journal of Research in Science Teaching*, 00: 1-24.
- Marzano, R. (2003). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Marzano, R., Pickering, D., and Pollack, J. (2001). *Classroom instruction that works: Research based strategies for increasing student achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Marx, R. W., Blumenfeld, P. C., Krajcik, J. S., Fishman, B., Soloway, E., & Geier, R., et al. (2004). Inquiry-based science in the middle grades: Assessment of learning in urban systemic reform. *Journal of Research in Science Teaching*, 41(10), 1063-1080.
- National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform*. Washington, DC: US Department of Education.
- National Committee on Science Education Standards and Assessment, National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press. from <http://www.nap.edu/catalog/4962.html>

- National Research Council. (2005). *How students learn: History, mathematics, and science in the classroom*. Washington, DC: The National Academies Press. from <http://www.nap.edu/catalog/10126.html>
- Porter, A. C., Garet, M., S., Desimone, L., Yoon, K. S., & Birman, B. F. (2000). *Does professional development change teaching practice? results from a three-year study* (Report to Department of Education, Office of the Under Secretary, Washington, DC. Washington, DC: American Institutes for Research in the Behavioral Sciences. (Eric Document Service No. ED455227) Retrieved February 17, 2005, from ERIC database.
- Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods, Second Edition*. Thousand Oaks, CA: Sage Publications, Inc.
- Ruiz-Primo, M. A. (2005). A multi-method and multi-source approach for studying fidelity of implementation. *Paper Presented at the Annual Meeting of the American Educational Research Association*, Montreal, Canada.
- 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.
- Scantlebury, K., Boone, W., Kahle, J. B., & Fraser, B. J. (2001). Design, validation, and use of an evaluation instrument for monitoring systemic reform. *Journal of Research in Science Teaching*, 38(6), 646-662.
- Schneider, R. M., Krajcik, J., & Blumenfeld, P. (2005). Enacting reform-based science materials: The range of teacher enactments in reform classrooms. *Journal of Research in Science Teaching*, 42(3), 283-312.
- Schwartz, R. S., Lederman, N. G., Khishfe, R., Lederman, J. S., Matthews, L., & Liu, S. (2002). *Explicit/reflective instructional attention to nature of science and scientific inquiry: Impact on student learning*(ED465622)
- Shadish, W.R. & Luellen, J.K. (2005). Quasi-experimental designs. In B. Everitt and D.Howell (Eds.) *Encyclopedia of behavioral statistics* (Volume 3, pp. 1641-1644). NY: Wiley.
- Shymansky, J. A., Hedges, L. V., & Woodworth, G. (1990). A reassessment of the effects of inquiry-based science curricula of the 60's on student performance. *Journal of Research in Science Teaching*, 27(2), 127-144.
- Shymansky, J. A., Yore, L. D., & Anderson, J. O. (2004). Impact of a school district's science reform effort on the achievement and attitudes of third- and fourth-grade students. *Journal of Research in Science Teaching*, 41(8), 771-790.

- Supovitz, J. A., & Turner, H. M. (2000). The effects of professional development on science teaching practices and classroom culture. *Journal of Research in Science Teaching*, 37(9), 963-980.
- Tal, T., Krajcik, J. S., & Blumenfeld, P. C. (2006). Urban schools' teachers enacting project-based science. *Journal of Research in Science Teaching*, 43(7), 722-745.
- Von Secker, C. E. (2002). Effects of inquiry-based teacher practices on science excellence and equity. *Journal of Educational Research*, 95(3 Jan-Feb), 151-160.
- Von Secker, C. E., & Lissitz, R. W. (1999). Estimating the impact of instructional practices on student achievement in science. *Journal of Research in Science Teaching*, 36(10), 1110-1126.
- Young, B. J., & Lee, S. K. (2005). The effects of a kit-based science curriculum and intensive science professional development on elementary student science achievement. *Journal of Science Education and Technology*, 14(5/6 December), 471-481.

**Table 1**  
Means and standard deviations of achievement test scores in Study 1

<b>Scaffolded Guided-Inquiry + Kits (Treatment)</b>				
<b>Measure</b>	<b>N Teachers</b>	<b>N Students</b>	<b>Mean</b>	<b>(SD)</b>
California Standards Test	10	276	6.030	2.086
FOSS Unit Post-test	10	276	16.800	4.759
FOSS Unit Pre-test	10	276	10.470	4.212
Gain	10	276	6.337	--

<b>Textbook-based Instruction (Control)</b>				
<b>Measure</b>	<b>N Teachers</b>	<b>N Students</b>	<b>Mean</b>	<b>(SD)</b>
California Standards Test	10	287	3.410	1.599
FOSS Unit Post-test	10	287	11.340	4.824
FOSS Unit Pre-test	10	287	10.440	4.950
Gain	10	287	0.899	--

**Table 2**  
Results of HLM analysis of pretest equivalence: Study 1

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p- value</b>
Intercept	10.464	0.545	19.200	< 0.001
SGI (Treatment)	-0.025	0.773	-0.032	0.975

**Table 3**  
Results of HLM analysis of treatment effects on the FOSS Post-Test:  
Study 1

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p-value</b>
Intercept	11.217	0.798	14.056	< 0.001
SIG (Treatment)	5.246	0.956	5.486	< 0.001
Pretest	0.823	0.023	35.948	< 0.001

**Table 4**  
Results of HLM analysis of treatment effects on the California Standards Test:  
Study 1

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p-value</b>
Intercept	3.373	0.315	9.590	< 0.001
SIG (Treatment)	2.580	0.452	5.708	< 0.001
Pretest	0.263	0.010	26.000	< 0.001

**Table 5**  
Means and standard deviations of achievement test scores in Study 2

<b>Scaffolded Guided-Inquiry + Kits (Treatment)</b>				
<b>Measure</b>	<b>N Teachers</b>	<b>N Students</b>	<b>Mean</b>	<b>(SD)</b>
California Standards Test	12	392	1.853	6.010
FOSS Unit Post-test	12	392	4.224	16.740
FOSS Unit Pre-test	12	392	3.377	10.790
Gain	12	392	3.922	5.954

<b>FOSS Kits Only (Control)</b>				
<b>Measure</b>	<b>N Teachers</b>	<b>N Students</b>	<b>Mean</b>	<b>(SD)</b>
California Standards Test	12	370	3.890	1.964
FOSS Unit Post-test	12	370	12.690	3.581
FOSS Unit Pre-test	12	370	11.060	3.331
Gain	12	370	1.624	2.764

**Table 6**  
Results of HLM analysis of pretest equivalence: Study 2

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p- value</b>
Intercept	11.217	0.619	18.107	< 0.001
SGI (Treatment)	-0.475	0.607	-0.782	0.443

**Table 7**  
Results of HLM analysis of treatment effects on the FOSS Post-Test:  
Study 2

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p-value</b>
Intercept	12.783	0.632	20.211	< 0.001
SIG (Treatment)	4.093	0.746	5.489	< 0.001
Pretest	0.643	0.038	17.022	< 0.001

**Table 8**  
Results of HLM analysis of treatment effects on the California Standards Test:  
Study 2

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p-value</b>
Intercept	3.985	0.309	12.893	< 0.001
SIG (Treatment)	2.169	0.358	6.061	< 0.001
Pretest	0.231	0.019	11.587	< 0.001

**Table 9**  
Means and standard deviations of achievement test scores in Study 3

<b>FOSS Kits Only (Treatment)</b> (Control Group in Study 2)				
<b>Measure</b>	<b>N Teachers</b>	<b>N Students</b>	<b>Mean</b>	<b>(SD)</b>
California Standards Test	12	370	3.890	1.964
FOSS Unit Post-test	12	370	12.690	3.581
FOSS Unit Pre-test	12	370	11.060	3.331
Gain	12	370	1.624	2.764

<b>Textbook Based (Control)</b> (Control Group in Study 1)				
<b>Measure</b>	<b>N Teachers</b>	<b>N Students</b>	<b>Mean</b>	<b>(SD)</b>
California Standards Test	10	287	3.410	1.599
FOSS Unit Post-test	10	287	11.340	4.824
FOSS Unit Pre-test	10	287	10.440	4.950
Gain	10	287	0.899	1.439

**Table 10**  
Results of HLM analysis of pretest equivalence: Study 3

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p- value</b>
Intercept	10.460	0.458	22.820	< 0.001
SGI (Treatment)	0.584	0.615	0.951	0.354

**Table 11**  
Results of HLM analysis of treatment effects on the FOSS Post-Test:  
Study 3

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p-value</b>
Intercept	11.062	0.484	22.866	< 0.001
SIG (Treatment)	1.706	0.546	3.123	0.006
Pretest	0.879	0.022	40.627	< 0.001

**Table 12**  
Results of HLM analysis of treatment effects on the California Standards Test:  
Study 3

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p-value</b>
Intercept	3.454	0.364	9.495	< 0.001
SIG (Treatment)	0.573	0.281	2.038	0.055
Pretest	0.271	0.013	21.608	< 0.001

**Table 13**  
Means and standard deviations of achievement test scores in Study 4

<b>Experienced SGI + Kits (Treatment)</b> (Treatment Group in Study 2)				
<b>Measure</b>	<b>N Teachers</b>	<b>N Students</b>	<b>Mean</b>	<b>(SD)</b>
California Standards Test	12	392	6.010	1.853
FOSS Unit Post-test	12	392	16.740	4.224
FOSS Unit Pre-test	12	392	10.790	3.377
Gain	12	392	5.954	3.922

<b>Novice SGI + Kits (Control)</b> (Treatment Group in Study 1)				
<b>Measure</b>	<b>N Teachers</b>	<b>N Students</b>	<b>Mean</b>	<b>(SD)</b>
California Standards Test	10	276	6.030	2.086
FOSS Unit Post-test	10	276	16.800	4.759
FOSS Unit Pre-test	10	276	10.470	4.212
Gain	10	276	6.337	4.397

**Table 14**  
Results of HLM analysis of pretest equivalence: Study 4

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p- value</b>
Intercept	10.434	0.640	16.316	< 0.001
SGI (Treatment)	0.284	0.862	0.330	0.745

**Table 15**  
Results of HLM analysis of treatment effects on the FOSS Post-Test:  
Study 4

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p-value</b>
Intercept	16.259	1.012	16.073	< 0.001
SIG (Treatment)	0.578	1.271	0.455	0.654
Pretest	0.575	0.038	15.180	< 0.001

**Table 16**  
Results of HLM analysis of treatment effects on the California Standards Test:  
Study 4

<b>Parameter</b>	<b>Estimate</b>	<b>SE</b>	<b>t ratio</b>	<b>p-value</b>
Intercept	5.744	0.440	13.052	< 0.001
SIG (Treatment)	0.266	0.537	0.496	0.625
Pretest	0.223	0.018	12.052	< 0.001