Effective Programs for Elementary Science: A Best-Evidence Synthesis

Robert E. Slavin
Johns Hopkins University
-and-
University of York

Cynthia Lake
Johns Hopkins University

Pam Hanley
University of York

Allen Thurston
Durham University

May, 2012

This research was supported by a grant from the National Science Foundation (No. DRL-1019306). However, any opinions expressed are those of the authors and do not represent NSF positions or policies.

Abstract

This article presents a systematic review of research on the achievement outcomes of all types of approaches to teaching science in elementary schools. Study inclusion criteria included use of randomized or matched control groups, a study duration of at least 4 weeks, and use of achievement measures independent of the experimental treatment. A total of 17 studies met these criteria. Among studies evaluating inquiry-based teaching approaches, programs that used science kits did not show positive outcomes on science achievement measures (weighted ES=+0.02 in 4 studies), but inquiry-based programs that emphasized professional development but not kits did show positive outcomes (weighted ES=+0.30 in 8 studies). Technology approaches integrating video and computer resources with teaching and cooperative learning showed promise (ES=+0.37 in 5 studies). The review concludes that science teaching methods focused on enhancing teachers' classroom instruction throughout the year, such as cooperative learning and science-reading integration, as well as approaches that give teachers technology tools to enhance instruction, have significant potential to improve science learning.

Effective Programs for Elementary Science:
A Best-Evidence Synthesis

The success of all students in science has become a priority in countries throughout the world, as governments have increasingly realized that their economic futures depend on a workforce that is capable in science, mathematics, and engineering (Kilpatrick & Quinn, 2009; Duschl, Schweingruber, & Shouse, 2007). A particular focus in policy discussions is on science in the elementary grades, where children's early attitudes and orientations are formed. Yet science education is particularly problematic in elementary schools. Numerous surveys have found that elementary teachers are often unsure of themselves in science, with little confidence in their science knowledge or pedagogy (Harlen & Qualter, 2008; Cobern & Loving, 2002; Pell & Jarvis, 2003). Since the appearance of the National Science Education Standards (National Research Council, 1996, 2000, 2012) and the recent National Research Council (2012) frameworks, there has been general agreement in the U.S. about what students should learn in science, and a consensus that science should be taught using inquiry-oriented methods that emphasize conceptual understanding rather than just facts. Yet beyond this broad agreement, what do we know about what works in elementary science? While there have been several reviews of research on various aspects of science teaching, there has not been a comprehensive review of evaluations of alternative approaches to elementary science education.

There have been several reviews of research on various aspects of science education, such as inquiry teaching (Anderson, 2002; Bennett, Lubben, & Hogarth, 2006; Minner, Levy, & Century, 2010; Shymansky, Hedges, & Woodworth, 1990), small-group methods (Bennett, Lubben, Hogarth, & Campbell, 2004; Lazarowitz & Hertz-Lazarowitz, 1998), and overall methods (Fortus, 2008; Hipkins et al., 2002; Schroeder, Scott, Tolson, Huang, & Lee, 2007). Yet the studies reviewed in all of these are overwhelmingly secondary, not elementary. For example, the Schroeder et al. (2007) review identified 61 qualifying studies, of which only 6 took place in elementary schools. Minner, Levy, & Century (2010), in a review of inquiry-based science instruction, found 41 of 138 studies to focus on elementary science, but many of these were of low methodological quality. The only review of all research on elementary science within the past 25 years is an unpublished bibliography of research and opinion about science education written for Alberta (Canada) school leaders (Gustafson, MacDonald, & d'Entremont, 2007). Further, experiments evaluating practical applications of alternative science programs and practices are rare at all grade levels. Vitale, Romance, & Crawley (2010), for example, reported that experimental studies with student learning as an outcome accounted for only 16% of studies published in the *Journal of Research in Science Teaching* in 2005-2009, and this percentage has declined since the 1980s. Most of the few experiments are brief laboratory-type studies, not evaluations of practical programs.

Review Methods

The review methods for elementary science applied in this paper are similar to those used in math by Slavin & Lake (2008) and Slavin, Lake, & Groff (2009), and in reading by Slavin, Lake, Chambers, Cheung, & Davis (2009). These reviews used an adaptation of a technique called best evidence synthesis (Slavin, 2008), which seeks to apply consistent, well-justified

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education's Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

standards to identify unbiased, meaningful information from experimental studies, discuss each study in some detail, and pool effect sizes across studies in substantively justified categories. Best-evidence syntheses are similar to meta-analyses (Cooper, 1998; Lipsey & Wilson, 2001), adding an emphasis on narrative description of each study's contribution and limiting the review to studies meeting the established criteria. They are also similar to the methods used by the What Works Clearinghouse (2009).

Literature Search Procedures

A broad literature search was carried out in an attempt to locate every study that could possibly meet the inclusion requirements. Electronic searches were made of educational databases (JSTOR, ERIC, EBSCO, Psych INFO, Dissertation Abstracts) using different combinations of key words (for example, "elementary students" and "science achievement") and the years 1980-2011. Results were then narrowed by subject area (for example, "educational software," "science achievement," "instructional strategies"). In addition to looking for studies by key terms and subject area, we conducted searches by program name. Web-based repositories and education publishers' websites were examined. We contacted producers and developers of elementary science programs to check whether they knew of studies we might have missed. Citations from other reviews of science programs, including all of those listed above, as well as studies cited in primary research, were obtained and investigated. We conducted searches of recent tables of contents of key journals, such as *International Journal of Science Education, Science Education, Journal of Research in Science Teaching, Review of Educational Research, Elementary School Journal, American Educational Research Journal, British Journal of Educational Psychology, Journal of Educational Research, Journal of Educational Psychology,* and *Learning and Instruction.* Articles from any published or unpublished source that meet the inclusion standards were examined, but these leading journals were exhaustively searched as a starting point for the review. Studies that met an initial screen for germaneness (i.e., they involved elementary science) and basic methodological characteristics (i.e., they had a well-matched control group and a duration of at least 4 weeks) were independently read and coded by at least two researchers. Any disagreements in coding were resolved by discussion, and additional researchers were asked to read any articles on which there remained disagreements.

Effect Sizes

In general, effect sizes were computed as the difference between experimental and control posttests (at the individual student level) after adjustment for pretests and other covariates, divided by the unadjusted posttest control group standard deviation. If the control group SD was not available, a pooled SD was used. Procedures described by Lipsey & Wilson (2001) and Sedlmeier & Gigerenzor (1989) were used to estimate effect sizes when unadjusted standard deviations were not available, as when the only standard deviation presented was already adjusted for covariates or when only gain score SD's were available.

Effect sizes were pooled across studies for each program and for various categories of programs. This pooling used means weighted by the final sample sizes, using methods described by Slavin (2008). The reason for using weighted means is to recognize the greater strength, stability, and external validity of large studies, as previous reviews have found that small studies tend to overstate effect sizes (see Rothstein, Sutton, & Borenstein, 2005; Slavin, 2008; Slavin & Smith, 2009).

Criteria for Inclusion

Criteria for inclusion of studies in this review were as follows.

1. The studies evaluated programs and practices used in elementary science, and were published in 1980 or later. Studies could have taken place in any country, but the report had to be available in English.

2. The studies involved approaches that began when children were in grades K-5, plus sixth graders if they were in elementary schools.

3. The studies compared children taught in classes using a given science program or practice with those in control classes using an alternative program or standard methods.

4. The program or practice had to be one that could, in principle, be used in ordinary science classes (i.e., it did not depend on conditions unique to the experiment).

5. Random assignment or matching with appropriate adjustments for any pretest differences (e.g., analyses of covariance) had to be used. Studies without control groups, such as pre-post comparisons and comparisons to “expected” scores, were excluded.

6. Pretest data had to be provided, unless studies used random assignment of at least 30 units (individuals, classes, or schools) and there were no indications of initial inequality. If science pretests were not available, standardized reading or math tests, given at pretest or contemporaneously, were accepted as covariates to control for initial differences in overall academic performance. Studies with pretest differences of more than 50% of a standard deviation were excluded because, even with analyses of covariance, large pretest differences cannot be adequately controlled for, as underlying distributions may be fundamentally different (Shadish, Cook, & Campbell, 2002). Studies using pretests with indications of ceiling or floor effects were excluded.

7. The dependent measures included quantitative measures of science performance. Experimenter-made measures were accepted if they covered content taught in control as well as experimental groups, but measures of science objectives inherent to the program (and unlikely to be emphasized in control groups) were excluded, for reasons discussed in the following section.

8. A minimum study duration of 4 weeks was required. This is much shorter than the 12-week minimum used in the Slavin & Lake (2008) math review and the Slavin et al, (2009) reading review. A rationale for this appears in the following section.

9. Studies had to have at least two teachers and 15 students in each treatment group. This criterion reduced the risk of teacher effects in single-teacher/class studies.

Methodological Issues Characteristic of Science Education Studies

Research on programs and practices in science education is characterized by several features that are important to consider in a review. Perhaps the most important of these is that many experimental studies of science programs and practices use measures designed by the researcher that are intended to assess content taught in the experimental group but not emphasized or taught at all in the control group. As one example, Vosniadou et al. (2001) evaluated an approach to teaching fifth and sixth graders about forces, energy, and mechanics. The control group received three weeks of ordinary instruction in mechanics, while the experimental group received an intensive program over the same period. The pre- and posttest, made by the experimenters, focused on the precise topics and concepts emphasized in the experimental group. The control group made no gain at all on this test from pre- to posttest, while the experimental group did gain significantly.

Were the students better off as a result of the treatment, or did they simply learn about topics that would not otherwise have been taught? It may be valid to argue that the content learned by the experimental group is more valuable than that learned by the control group, but the experiment does not provide evidence that this particular experimental approach is better than traditional teaching, as the outcomes could be simply due to the fact that the experimental group was exposed to content the control group never saw. A study reported by Slavin & Madden (2011), focusing on math and reading studies reviewed in the U.S. Department of Education's What Works Clearinghouse (WWC), found that such measures that are "inherent" to the treatment are associated with effect sizes that are much higher than are measures of the curriculum taught in experimental as well as control groups. For example, among seven mathematics studies included in the WWC and using both treatment-inherent and treatment-independent measures, the mean effect sizes were +0.45 and -0.03, respectively. Among ten reading studies, the mean effect sizes were +0.51 and +0.06, respectively. In science, experimenter-made measures inherent to the content taught only or principally in the experimental condition are often the only measures reported.

Another recent example of the problem of treatment-inherent measures is a study by Heller et al. (2012) comparing three professional development strategies for teaching fourth graders a unit on electric circuits. Students were pretested and then posttested on a test "…designed to measure a *Making Sense of SCIENCE* content framework…" (Heller et al., 2012, p. 344). The three experimental groups all implemented the *Making Sense of SCIENCE* curriculum unit on electric circuits, while the control teachers may not have even been teaching electric circuits during the same time period and certainly could not be assumed to be teaching the same content contained in the *Making Sense of SCIENCE* curriculum. (The only indication that they were teaching electric circuits at any point in fourth grade was a suggestion that this topic typically appears in fourth grade standards, but even if control teachers did teach electric circuits, they may have done so before or after the experimental period.) Comparisons among the three experimental conditions in this study are meaningful, but the comparisons with the control group are not, because comparisons with the control group may simply reflect the fact that experimental teachers were teaching about electric circuits during the experimental period and control teachers were not doing so.

The issue of treatment-inherent vs. treatment-independent measures is related to that of proximal vs. distal measures (see, for example, Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002), but it is not the same. A proximal measure is one that closely parallels an enacted curriculum, while a distal measure is, for example, a state assessment or standardized test. Not surprisingly, students in experimental treatments generally show more gain over time on proximal than on distal measures, as was found in the Ruiz-Primo et al. (2002) study. However, in a study involving a comparison with a control group rather than just a pre-post gain, the question is whether the control group was exposed to the assessed content. A proximal measure in such a study would be meaningful if the content it assesses was also taught to the control group (using a contrasting method), but even the most "distal" measure is not useful in a control group comparison if the content of that measure was not taught to the control group. The question in a control group comparison study is whether a measure is "fair" to both groups, not whether it is proximal or distal.

Another issue of particular importance in science is the duration of the study. In prior reviews of math and reading, we have used a duration of 12 weeks as an inclusion criterion, on the basis that shorter studies often create unusual conditions that could not be maintained over a full year. For example, the Vosniadou et al. (2001) study of force and energy was able to provide extraordinary resources and classroom assistance to the experimental classes over a 3-week period. This is perhaps justifiable for theory building, but one might question whether principals or teachers should select programs or practices based on their evidence of effectiveness in such brief and artificial experiments, since instructional resources and arrangements were provided that could not be maintained for a significant period of time. Because science studies often focus on a single well-defined topic, such as cell functions or electricity, for a few weeks, we reduced our duration requirement to four weeks, but such brief experiments should be interpreted with caution.

A study by Baines, Blatchford, & Chowne (2007) provides an internal comparison that illustrates the problems of brief experiments. The study contained both a year-long evaluation of a cooperative learning intervention with an appropriately broad measure, and a brief, embedded experiment with a measure closely aligned to the experimental content. The overall evaluation, described in more detail later in this paper, found modest positive effects (ES=+0.21, p<.01) in comparison to a matched control group. However, a two-week "micro-study" on evaporation found much more positive outcomes: ES=+0.58 (p<.001). Even on the end-of-year test, items on evaporation and on another micro-unit on forces (whose short-term outcomes were not reported) accounted for the whole program effect. Effect sizes for these items were +0.43 for evaporation (p<.001) and +0.29 for forces (p<.001), but on the remaining items, the effect size was +0.09 (n.s.). Presumably, experimental teachers focused substantially more time and energy on these micro-units than on the rest of the curriculum, even though control teachers were also supposed to have been teaching evaporation and forces during the same time periods. This study demonstrates why brief experiments with targeted measures cannot be used as indicators of the practical effectiveness of science programs or practices. Had the "micro-unit" experiment been the only one reported by Baines et al. (2007), it would have given substantially inflated indicators of the effectiveness of the treatment.

A design feature often seen in evaluations of science programs, especially those with brief durations and small sample sizes, is the provision of extraordinary resources or clearly non-replicable conditions to experimental classes. For example, a study of mastery learning by Arlin & Webster (1983) taught children about sailing. After four one-hour lessons, children in the experimental group were given a formative test. Those who scored less than 80% were given 1-1 tutoring for up to four more hours, doubling their instruction time. Control students did not receive tutoring. Impacts at the end of the 8-hour study were substantial, with an effect size of +3.0! However, such a study has little meaning for practice, since providing four hours of one-to-one tutoring added to four hours of group instruction is impractical, and since the topic, sailing, was chosen to require no prerequisite skills or knowledge, an unusual situation in real classroom conditions. Brief studies of this kind, intended to add to theory rather than practice, often create artificial conditions and provide additional staff to work with children.

As another example, a 21-day study of cooperative learning by Johnson, Johnson, Scott, & Ramolae (1985) compared students who worked in small groups to those who worked on their own in an “individualistic” condition. Both groups studied a written unit on electricity. However, there was no teaching provided to either group. This meant that children in the cooperative condition were likely to have peers who could explain concepts to them. Those in the “individualistic” group received no teaching and were forbidden to talk to their peers. Having students receive no teaching can be arranged for a 21-day experiment, but could not of course be maintained over a long period.

Interventions that used expensive technology or a great deal of professional development and coaching were included in the review, on the basis that such investments might be justified if outcomes were very positive or if the interventions might become less expensive over time. However, we excluded studies evaluating interventions that appeared to depend on providing children with such extraordinary human resources (such as graduate students working in experimental but not control classes every day the experiment is in operation) that clearly could not be provided over extended time periods. The 4-week duration requirement excluded most such artificial experiments, but in just a few cases longer experiments that provided extraordinary resources or non-replicable conditions to the experimental group were excluded.

Findings

The most important finding of the present review is the very limited number of rigorous experimental evaluations of elementary science programs. After an exhaustive search involving examination of 327 published and unpublished articles that purported to evaluate science approaches in elementary schools since 1980, only 17 studies met the review standards. Supplementary Table S- 1, Studies not Included in the Review, lists the studies that were considered but rejected, and the main reason for exclusion. As a point of comparison, a review of elementary mathematics programs using a somewhat more stringent set of inclusion standards (requiring a treatment duration of at least 12 weeks instead of 4) identified 87 qualifying studies (Slavin & Lake, 2008).

The few elementary science studies that did meet the inclusion criteria provide useful information on several approaches to improving outcomes in science teaching. Most of the qualifying studies focused on *inquiry-oriented instructional processes* for teachers, including approaches such as cooperative learning and integrating science and reading. The theory of action uniting this category of approaches is an emphasis on teachers improving science learning by using specific, well-articulated strategies designed to develop students' understanding, curiosity, and ability to apply scientific methods. These interventions invariably emphasize professional development and coaching to help teachers use promising approaches.

Two categories of inquiry-oriented instructional process programs were designated: Those that also provided teachers with kits and specific guidelines for hands-on inquiry-oriented explorations, and those that provided professional development without kits. The theory of action underlying the programs providing kits emphasizes the idea that if teachers have well-designed materials to enable them to teach inquiry lessons, as well as professional development to help them use these materials, they are more likely to effectively implement the programs, and student outcomes will improve. Examples of this approach are science kits such as FOSS (Full-Option Science System) and STC (Science and Technology for Children). These provide extensive professional development, but the main focus is on providing teachers with appealing, well-developed materials to help them use inquiry and laboratory approaches as well as traditional content.

Another category of approaches to improving science instruction emphasizes the use of *technology* to enhance student outcomes. This category includes individual technologies, such as computer-assisted instruction, as well as class-focused technology, such as video and interactive whiteboard technologies, and combinations of these types.

Inquiry-Oriented Instructional Process Programs Without Science Kits

Inquiry-oriented instructional process programs that do not provide specific materials focus their efforts on helping teachers learn and use generic processes in their daily science teaching, such as cooperative learning, concept development, and science-reading integration. Table 1 summarizes characteristics and findings of the eight qualifying studies of interventions in this category.

=================

TABLE 1 HERE

=================

Increasing Conceptual Challenge. Mant, Wilson, & Coates (2007) evaluated a professional development program in 32 mostly rural and village schools in Oxfordshire, England. Almost all children were White, and few qualified for free school meals. Teachers of Year 6 (ages 10-11) in 16 schools were provided with extensive professional development intended to increase engagement and conceptual challenge in science lessons. Sixteen control schools were matched on prior scores on the national science exam (number of students receiving scores of "5," the top score), number of children in Year 6, and percent of students with special needs.

In each experimental school, the science coordinator and a Year 6 class teacher participated in an extensive series of professional development sessions, consisting of 8 full-day and 4 evening trainings at Oxford Brookes University. The sessions emphasized cognitively challenging, practical, whole-class science lessons. Teachers learned to use thinking skills strategies such as regular "bright ideas time" opportunities for focused discussion, "positive, minus, and interesting" (PMI) features of phenomena, and "big questions." Teachers were encouraged to emphasize higher-order thinking, practical work, investigations, and purposeful, focused recording. The content and materials used in experimental and control schools was the same, as dictated by the National Curriculum for England.

The evaluation compared Key Stage 2 science tests routinely administered to all students in England at the end of elementary school (Year 6). Students' tests are rated on a scale from 1 to 5, with 4 considered passing and 5 outstanding. The year before the experiment, experimental and control schools were nearly identical in percent of students attaining Level 5 (E=39.6%, C=39.4%). At the end of the study year, however, 51.4% of experimental students and 41.6% of control students reached level 5. This difference was statistically significant at the school level ($p<.05$), and was equivalent to an individual-level effect size of +0.33, with estimated N's for each condition of E=560, C=560.

Science IDEAS. A study by Baines, Blatchford, & Chowne (2007) evaluated a cooperative learning intervention in 21 classes in 12 London elementary schools (N=560). Students were in Years 4-5 (8-10 years old). Control students were in 40 classes in 19 schools (N=1027) in a different area of London. The schools were selected in the year following the experimental year to match the experimental schools in demographics and pretests.

The cooperative intervention, called *Social Pedagogical Research in Group Work (SPRinG)*, involved students working in groups of 2-4 on a regular basis over the course of a year. Teachers participated in 7 half-day meetings, and were given manuals and lesson plans to provide a structure and examples of cooperative work. Students were trained in cooperative skills such as listening, explaining, and sharing ideas, and these skills were reinforced during implementation.

Pre- and posttests were constructed from items adapted from standardized tests for Year 6, simplified for younger children. They included both multiple choice and open-ended items and emphasized interpretation of diagrams, tables, and graphs. Controlling for pretests, the overall effect size was +0.21 ($p<.01$).

As noted previously, embedded within the overall experiment was a “micro-experiment” in which students in experimental and control groups were pre- and posttested on a unit on evaporation, and then on a unit on forces. As noted earlier in this article, an evaluation of the two-week evaporation unit produced much larger effect sizes than those reported for the whole year, but did not meet the duration standards of this review. It is interesting to note that on the end-of-year tests as well, outcomes for questions relating to evaporation and forces had very positive outcomes, and analyses of the items other than evaporation and forces showed no experimental-control differences, suggesting that teachers emphasized these topics much more in the experimental than in the control group.

A concern frequently expressed by science educators is that in elementary schools driven by math and reading tests, science is often pushed aside. An approach developed by Romance & Vitale (2011) confronts this problem with a program called *Science IDEAS*, which integrates science with reading and focuses on building content-area reading skills as well as science skills. In a study reported in 2011, teachers in the experimental group received extensive professional development and coaching to help them build comprehension strategies for science and to build science concepts. Students were taught to link together observed events, to make productions or manipulate conditions to produce outcomes, and to make meaningful interpretations of events. The science approach emphasized hands-on activities, concept mapping, and journal writing. In particular, students were taught to read and to create propositional concept maps to represent scientific phenomena. Schools adopt *Science IDEAS* throughout the school and use it every day in a 1 ½ to 2-hour science/reading block. Project staff regularly visited teachers to monitor fidelity of implementation.

A six-year longitudinal evaluation of *Science IDEAS* was carried out in elementary schools in grades 3-8 in a large urban district in Florida. 12 schools implementing Science IDEAS were matched with 12 similar schools that served as a control group. The control schools used the district basal series in reading and the district science curriculum in a separate daily half-hour program. Overall, 40% of students received free lunches, 47% were White, 29% African-American, 19% Hispanic, and 5% Other. Controlling for pretests, analyses showed significantly higher scores for students at the end of eighth grade in Science IDEAS than for those in the control both for ITBS Science (ES= +0.29, p<.008) as well as for ITBS Reading Comprehension (ES= +0.40, p<.004).

Romance & Vitale (1992) evaluated an earlier version of Science IDEAS in a program that replaced a district-adopted reading textbook approach with a program that integrated reading with science, introducing content-area-reading strategies, hands-on science activities, and science process skills. Students in the experimental group participated in a combined daily 2-hour reading/science block, while those in the control group maintained a 1 ½ hour reading/language arts block, using the district’s basal series and a ½ hour science period, mostly using the district science text. Because of the limited time allocated to science in the control classes, teachers had fewer opportunities to use hands-on activities or to pursue science topics in depth.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education’s Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

The evaluation compared 3 fourth-grade classes (N=51) using the experimental program to 4 control classes (N=77) in a demographically similar school with similar pretest scores, all located in a large urban district in Florida. The treatments were implemented over a school year.

On routinely administered standardized tests, students in the experimental group scored substantially higher than controls on MAT-Science (ES=+0.90, p<.001) and also on ITBS Reading (ES=+0.40, p<.01). The science difference amounted to almost a full grade equivalent, while the reading difference was about 25% of a grade equivalent.

Collaborative Concept Mapping. Jang (2010) reported an evaluation of a collaborative concept-mapping technique in fourth-grade science classes in Taiwan. In the experimental classes, two teachers worked together as a team. Students (N=58) worked in small groups on activities that emphasized creating concept maps to organize information and ideas. Students discussed together, but then made their own learning journals and concept maps. The experiment compared two experimental to two control classes in an 8-week study focusing on electricity and rainbows. The matched control classes (N=56) received whole-class instruction using the same materials and activities, but without team teaching or team learning. The outcome measure was a schoolwide uniform science test ordinarily given by the schools. Adjusting for pretests, posttest scores significantly favored the experimental group (ES=+0.54, p<.05).

Systematic Vocabulary Instruction. Rosebrock (2007) evaluated a method for teaching science vocabulary in a middle-class suburb of Houston. The method taught fifth graders 35 terms relating to Earth and space science over a period of 12 weeks. Each of the terms, selected from the Texas state science standards, was introduced in a 12-step process in which the words were introduced, defined, explained, read in various contexts, demonstrated in hands-on lab work, discussed in small groups, illustrated in writing, concept maps, or diagrams, used in games and crossword puzzles, and finally quizzed. The experiment compared one school that used the vocabulary intervention and one that served as a control group. The schools were well matched on state test scores but not on demographics; the experimental school had a greater number of African American students (20% vs. 10%), Hispanic students (20% vs. 17%), and Asian students (13% vs. 5%) and fewer White students (48% vs. 68%) than controls. Overall, 16% of students were economically disadvantaged. There were 401 students in the experimental group and 285 controls.

The posttest measure consisted of the nine multiple-choice items relating to Earth and space science on the 40-item Texas Assessment of Knowledge and Skills (TAKS) science test. The author was unaware of how much overlap there was between the 35 vocabulary words taken from the state standards and the nine TAKS items relating to Earth and space science. Controlling for TAKS pretests, students in the vocabulary intervention scored significantly higher than controls (ES= +0.24, p<.001).

TEAMS. Scott (2005) carried out a year-long matched evaluation of an extensive science professional development approach called *TEAMS* (*Teachers Engaged in Authentic Mentoring Strategies*). The program provided teachers with a two-week summer institute, professional development days, mentoring from a building science specialist, monthly after-school meetings, classroom observation days, and participation in an electronic database system. These resources were intended to help teachers learn and effectively implement inquiry approaches to science teaching that emphasized engagement, exploration, elaboration, and evaluation. Teachers were taught reading and vocabulary strategies as applied to reading science content. They learned to use formative assessments for science teaching.

The study took place in Aldine, Texas where the author was science director. Aldine is a large, diverse district outside of Houston. Although the *TEAMS* process was used throughout many elementary schools in Aldine, the study evaluated third graders taught by 3 teachers in only three experimental schools and three control teachers in three similar schools matched on pretests and demographic factors. The TEAMS schools averaged 83% free lunch and 40% Limited English Proficient. Fifty-four percent of students were Hispanic, 37% African American, and 5% White. ITBS-Science data were obtained at the end of second grade (pretest) and the end of third grade for a total of 66 experimental and 33 control students. Adjusting for pretest differences, posttest differences favored the TEAMS students (ES=+0.29).

4-E Learning Cycle. In a small study in Kuwait, Ebrahim (2004) evaluated a 4-E Learning Cycle in four fourth-grade classes. The experimental treatment emphasized exploration, explanation, expansion, and evaluation, using experiments, student-centered, cooperative work, assessment through teacher observations rather than student tests, and real-world problem solving. The control group used a traditional lecture format to cover the same content, a month-long unit on plants.

The study compared two classes in each treatment. Each of two teachers taught one 4-E and one control class (N= 49E, 49C). Because Kuwaiti classes are segregated by gender, there was one class of boys and one of girls in each treatment.

Students were pre- and posttested using an experimenter-made test of the content taught equally in both conditions. The groups were well-matched overall at pretest. At posttest, differences on the posttest strongly favored the 4-E groups (ES=+0.96, p<.001).

The overall weighted mean effect size for the 8 qualifying studies of inquiry-oriented professional development without science kits was +0.30.

Instructional Process Programs With Science Kits

Instructional process programs that provide teachers with specific materials and instructions resemble those discussed in the previous section in that they provide teachers with extensive initial training and coaching. However, they are different in focus, in that they tend to emphasize the rich content supported by their materials rather than focusing on all of science education. That is, the theory of action in science kit programs is that implementing the hands-on activities will build deep learning about the scientific process and about the core concepts of elementary science. There may be less of an emphasis on the direct teaching of science concepts that takes place during times when kits are not being used. This contrasted with the approaches described in the previous section, which tended to focus equally on generic strategies for inquiry and hands-on experiments and on strategies for concept development that applied to all science taught in the elementary grades.

Table 2 summarizes characteristics and findings of the four qualifying studies of instructional process approaches that provide specific student inquiry activities and materials.

===============

TABLE 2 HERE

===============

Insights, FOSS, and STC. Pine et al. (2006) carried out a major, large-scale evaluation of the impacts of the major hands-on inquiry curricula developed in the 1990's: *Insights*, *FOSS* (*Full-Option Science System*), and *STC* (*Science and Technology for Children*). The study compared fifth graders in 41 classrooms in 9 school districts in California, Arizona, and Nevada. Two groups of schools using hands-on inquiry curricula over the course of a year were identified: high-SES (less than 50% free lunch; mean=21%) and low-SES (more than 50% free lunch; mean=64%). Then matched schools using traditional textbooks were identified. Approximately 500 students were in each treatment condition. In order to control for any pre-existing differences, students were given a standardized Cognitive Abilities (CogAt) test, focusing on reading and math. This was given at about the same time as the outcome measures, so this is a contemporaneous control variable rather than a pretest. Two tests were used to assess outcomes. One was a 25-item selection of items from the Third International Math and Science Study (TIMSS), with 23 multiple-choice and 2 open-ended questions. The second was a performance measure developed by the investigators. Students were asked to carry out four experiments, one involving determining weight using a spring, one testing the absorbency of different paper towels, one comparing melting rates of ice cubes in salt vs. fresh water, and one involving observations of flatworms over 3 days. Each of the performance measures, administered one-on-one by research assistants, yielded scores on planning an inquiry, observation, data collection, graphical and pictorial representation, inference, and explanation based on evidence.

A total of 720 students took all measures. After adjustments for the CogAt, there were no differences between inquiry and textbook students on the TIMSS items (mean ES= -0.02). There were significant differences favoring the inquiry students on the flatworms task ($p<.05$), but not on the other measures. Averaging across the four performance measures, the mean effect size was +0.11. An HLM analysis, with students nested within classrooms, also found a small positive effect for the flatworm task, but no significant differences for the four tasks taken together. There were no interactions with gender or socioeconomic status.

A 14-week study involving fifth graders was carried out by Leach (1992) in an urban district in Texas with a minority enrolment of 49%. Students were randomly assigned to one of two experimental and three control classes (N=38E, 65C). Control classes were taught three chapters from a textbook, while experimental students used three FOSS units. The only overlap in content was a unit (*FOSS*) and chapter (control) on electricity and magnetism. The experimenter selected items on this topic from the control group's textbook for use as a posttest, and CTBS science was also used as a posttest. On CTBS, effects non-significantly favored the *FOSS* students (ES= +0.29, n.s.). On the electricity and magnetism test, effects were statistically significant and much larger (ES= +0.67, p<.02). However, it was unclear whether the amount of time and focus on electricity and magnetism was similar in the two conditions.

Young & Lee (2005) reported a study of outcomes of use of *FOSS* and *STC* science kits that does not meet the standards of this review, but is worthy of a brief note given the limited evidence base for these hands-on inquiry kits. The study compared fifth graders who attended schools that had used *FOSS* or *STC* for several years. The problem with the study is that it pretested students at the beginning of fifth grade, when most would have already had several years of experience with *FOSS* or *STC*. That is, if the hands-on inquiry kits were effective in these schools, the pretests might already reflect this. In fact, the experimental students in this study scored significantly higher than textbook controls at pretest (ES=+0.23, p<.02), and also at posttest (ES= +0.22, p<.03), meaning that there was no further relative gain. The pretest difference cannot, of course, be attributed to the use of the science kits, but it is interesting to know that there were no further gains during fifth grade. Because of the problem with pretesting after the treatments were already under way, the Young & Lee (2005) study does not appear in Table 2.

SCALE. G. Borman, Gamoran, & Bowdon (2008) evaluated a large-scale professional development initiative in the Los Angeles Unified School District (LAUSD). The intervention was a National Science Foundation Math and Science Partnership initiative, called *SCALE*, for *System-Wide Change for All Learners and Educators*. In the *SCALE* elementary science component, fourth- and fifth-grade teachers participated in summer institutes and then received coaching and mentoring in the use of extended, inquiry-based "immersion units" intended to take students and teachers through a full cycle of inquiry in science investigation. The units emphasized "big ideas," posing scientific questions, giving priority to evidence, connecting evidence-based explanations to scientific knowledge, and communicating and justifying explanations. One teacher in each grade level participated in the summer institute, but all teachers received extensive coaching and mentoring at their school.

Eighty schools were randomly assigned to experimental or control conditions. A few schools had missing data, and the analysis sample included 33 experimental and 38 control schools. Control schools were offered the *SCALE* curriculum units, but not the professional development or ongoing coaching. Approximately 73% of students were Hispanic, 11% were White, 8% were African-American, 3% were Asian, and 3% were Filipino. 76% of students received free lunch, and 33% were English language learners. Experimental and control schools were well matched on these factors and on reading and math scores.

During the first program year, the outcome measures were three science assessments provided by LAUSD to all students in grades 4-5. One test focused on life science, one on earth science, and one on physical science. Each consisted of 20 multiple choice items and one constructed-response item. Teachers were allowed to give these tests in any order.

Hierarchical linear modeling (HLM) was used to analyze the data, controlling for science pretests and other factors. On life science, the treatment effects were significantly negative (ES= -0.27, p<.01), while on earth science (ES= +0.01, n.s.) and physical science (ES= -0.08, n.s.) there were no differences, for an average effect size of -0.11. Additional analyses investigated these unexpected findings. One hypothesis was that effects might be more positive for the science lead teachers who actually participated in the summer training. However, the students of the lead teachers scored slightly worse, relative to controls, than did teachers in general. Another analysis found that for teachers in general, treatment effects were the same for experienced teachers (>3 years) than for less experienced teachers. However, students of lead teachers with less experience gained slightly from the *SCALE* treatment while students of more experienced lead teachers did worse than controls. Examinations of outcomes on life science questions more closely aligned to the *SCALE* curriculum did not show positive outcomes.

Gamoran, G. Borman, Bowden, Shewakramani, & Kelly (2012) followed students in the G. Borman et al. (2008) study for an additional year. At the end of that time, achievement results were no longer significantly negative, but they were essentially zero on all LAUSD measures: life science (ES= -0.05, n.s.), earth science (ES= +0.03, n.s.), and physical science (ES=-0.03, n.s.). On state standardized tests given to fifth graders, differences were also very small on life science (ES= -0.02, n.s.), earth science (ES= -0.02, n.s.), and physical science (ES = -0.03, n.s.).

Teaching SMART. Another large-scale, randomized evaluation of science kits was carried out by K. Borman, Boydston, Lee, Lanehart, & Cotner (2009). They evaluated the *Teaching SMART* professional development program in Pasco County, Florida. Twenty schools and their teachers of grades 3-5 were matched on pretests and demographic factors and then randomly assigned to *Teaching SMART* or control conditions (N (schools)=10E, 10C) over a three-year period. *Teaching SMART* professional development emphasized an exploratory, hands-on approach, cooperative learning, equity, questioning techniques, problem solving, discovery, and real-world applications. In addition to initial inservices, teachers received extensive on-site coaching from specially trained site coaches (each of three site coaches was responsible for about 40 teachers). The program provides more than 100 "culturally sensitive, grade-specific" lesson plans based on AAAS and NSF standards and benchmarks, as well as activity kits with consumable supplies and equipment kits with all necessary resources.

Student achievement was measured on the PASS (Partnership for the Assessment of Standards-based Science), which combined authentic performance assessments with multiple-choice items. PASS assessments were administered as pretests and then at the end of third, fourth, and fifth grades. Data from routinely administered state FCAT reading and math tests were also collected and reported.

Outcomes on the PASS over the 3-year experiment were not statistically or educationally significant. Adjusting for pretests, there was no significant difference on the PASS multiple choice items (ES=+0.08, n.s.), and no difference (ES= .00, n.s.) on the performance measures.

Overall, the weighted mean for programs using science kits was +0.02, which is effectively zero.

Technology Applications

Although there have been substantial investments made by NSF and other government and private funders throughout the world in development and evaluation of technology solutions for science education, only five studies of technology programs in elementary science met the standards of this review. The many articles on technology programs that did not meet the review standards typically described studies of very brief duration, often carried out under very artificial circumstances (e.g., with many additional staff members helping children with the technology). Perhaps most importantly, many studies of technology programs in science that did not qualify for this review used measures inherent to the experimental program and did not ensure that there was a control group studying the same content. It is interesting to note that in systematic reviews of research on elementary math (Slavin & Lake, 2008) and reading (Slavin et al., 2009), studies of technology programs, especially computer-assisted instruction, was the category with the largest number of qualifying studies. The inclusion standards in those reviews were nearly identical to those used in this review.

Table 3 summarizes characteristics and outcomes of the five studies of technology-focused programs that met the standards of the present review.

=================

TABLE 3 HERE

=================

BrainPOP. In an Israeli study, Barak, Ashkar, & Dori (2011) evaluated a program in which whole classes were shown on-line multimedia content called *BrainPOP* (http://www.brainpop.com). *BrainPOP* students viewed 3 to 5 minute animated *BrainPOP* videos that explain scientific concepts in an interesting way. A teacher's section provides lesson plans and ideas for building on the *BrainPOP* content. In this experiment, students saw about one video each week. They then engaged in activities either individually or in cooperative pairs, with teacher instruction following up on the concepts introduced in the videos. The *BrainPOP* videos and follow-up activities were organized to align with the Israeli national curriculum. Control classes used traditional textbooks and classroom teaching to study the same content, equally aligned with Israeli standards.

The experiment took place over the course of a school year. A total of 926 fourth and fifth graders in 5 elementary schools received the experimental treatment, while 409 students in two schools matched on pretests and parent characteristics served as a control group. Students were pre- and posttested on a measure of "understanding of scientific concepts and phenomenon," based on Israeli national standards. Adjusting for pretests, the posttest means strongly favored the experimental group (ES=+0.43, p<.001). Ratings of students' explanations also favored the experimental group (p<.05).

SEG Research (2009) carried out an evaluation of *BrainPOP* in Palm Beach County, Florida, and New York City. Third and fifth graders who used *BrainPOP* 2-3 hours per week (N=186) were pre- and posttested on Stanford-10 scales, and compared to matched control students (N=185). On the science scale, *BrainPOP* students gained significantly more than controls in fifth grade (ES=+0.55, p<.001) but not third grade (ES=+0.10, n.s.), for a mean of +0.33. Positive effects were also reported for SAT-10 measures of reading, vocabulary, and language for fifth graders, and for reading and vocabulary among third graders.

The Voyage of the Mimi. *The Voyage of the Mimi* (Bank Street, 1984) is a multimedia program that uses a variety of technology related to whales to teach science in elementary and middle schools. Rothman (2000) evaluated an application of the program in three schools in a Philadelphia suburb. At the time of the study, the program included computer simulations and modeling, microcomputer-based laboratory data collection and analysis, and interactive video disks that showed students appealing video content on the topics of study. In the Rothman (2000) evaluation, four modules were used: “Introduction to Computing,” “Maps and Navigation” (in which student teams use science and math to help free a whale caught in the net of a fishing trawler), “Ecosystems” (two computerized simulations in which students observe changes in populations of animals and plants as ecosystems change), and “Whales and Their Environment” (hands-on microcomputer activities in which students collect data about temperature, light, and sound to test hypotheses related to whales).

The study compared a total of 163 fifth graders in three schools. One implemented all four of the *Mimi* modules and participated in a *Mimi*-oriented field trip. In the four fifth-grade classes (n=57), the author estimated that *Mimi* activities were used 37% of class periods, leaving 63% for traditional textbook instruction. A second school with four classes (N=54) used only one *Mimi* module, for 7% of class periods, and a control school with three classes (n=52) only used the textbook.

Students were pre- and posttested on a 40-item Metropolitan Achievement Test (MAT-7) science scale in a year-long experiment. Students in the school that used the full program gained non-significantly more than the control school (ES=+0.25, n.s.), and the school that made minimal use of the program also gained non-significantly more than the control students (ES=+0.33, n.s.). On an attitude measure, only the full treatment school gained significantly more than the control school (p<.015).

Web-Based Labs. In a study in two Taiwan elementary schools, Sun, Lin, & Yu (2008) evaluated an approach in which fifth graders used web-based lab simulations to do experiments. Two 4-week units were taught, one on acids and alkalis and one on the operation of a microscope.

In each of several lab exercises, students were shown computer screens. On the left side, they carried out simulated experiments, while on the right side were “cabinets” containing simulated tools and instruments, such as thermometers, alcohol burners, and test tubes. Students could use the simulated equipment and see the results of their work; for example, moving a simulated magnet near a simulated compass would cause the needle to point toward the magnet.

The Best Evidence Encyclopedia is a free web site created by the Johns Hopkins University School of Education’s Center for Data-Driven Reform in Education (CDDRE) under funding from the Institute of Education Sciences, U.S. Department of Education.

Records of students' operations were made immediately available to the teacher, who could then respond right away to errors.

The experiment compared four intact classes in two schools. Classes were randomly assigned to experimental (N=56) or control (N=57) conditions, but with such a small number of classes the design was treated as matched. Control classes were taught precisely the same content as were experimental students, and the same amount of time was allocated to each group. Detailed lesson plans were given to each teacher to try to standardize the content taught in each treatment group.

Students were pre- and posttested on experimenter-made tests covering the content taught in all classes. Adjusting for (small) pretest differences, students using the web-based labs scored higher than controls (ES=+0.30, p<.05).

In a closely-related experiment, Sun, Lin, & Wang (2009) evaluated use of a 3-D virtual reality (VR) model of the sun and moon in a 4-week unit. Taiwanese fourth graders in the VR group used a unit called "Capricious Moon Lady" focusing on location of the moon, phases of the moon, relation of the moon phases to the lunar calendar, and related topics. The computer was able to simulate the positions of the Earth and moon, 3-D coordinates, effects of the gravitational pulls of sun, Earth, and moon, and so on. Students could choose to "observe" the sun, Earth, and moon from the Earth, from a movable space ship, or from a spaceship in a set orbit. They went through a series of exercises to learn about the moon's phases and movement, and also used the software to analyze their own observations of the moon each evening. Control students studied the same content, but used 2-D photographs to learn about the moon. Control students also observed the moon each evening, but did not enter their observations on the computer.

In four intact classrooms within an elementary school in southern Taiwan, students were selected from two treatment or two control classes in a matched design (T=63, C=65). Students were pretested and posttested on experimenter-made measures keyed to the content studied by both groups. At the end of the 4-week experiment, the treatment group scored significantly better, adjusting for small pretest differences (ES=+0.26, p<.02).

Across the five studies of technology applications, the weighted mean effect size was +0.37.

Discussion

As noted earlier, the most important findings of this review is the fact that very few studies of elementary science met the inclusion standards. Out of 327 identified studies purporting to evaluate science approaches in elementary schools, only 17 had control groups, durations of at least four weeks, equivalence on pretests, and measures not inherent to the experimental treatment. In light of the small numbers of qualifying studies, it must be acknowledged that any conclusions about the findings of these studies can only be tentative.

Previous syntheses of research on science teaching have reported much more positive impacts on science achievement than those found in the current synthesis. For example, a meta-analysis by Schroeder et al. (2007) reported mean effect sizes ranging from +0.29 to +1.28 for 8 categories of science treatments in elementary and secondary schools, far higher than those reported in the present review. However, Schoeder et al. (2007) included experiments using treatment-inherent measures, brief studies, and artificial procedures characteristically associated with high positive effect sizes.

A surprising finding from the largest and best-designed of the studies is the limited achievement impact of elementary science programs that provide teachers with kits to help them make regular use of hands-on, inquiry-oriented activities. These include evaluations of the well-regarded *FOSS*, *STC*, *Insights,* and *Teaching SMART* programs, none of which showed positive achievement impacts. One might argue that traditional science tests might not be sensitive to the more sophisticated understandings of scientific process that are the targets of these inquiry-oriented approaches, but the studies by Pine et al. (2006) and K. Borman et al. (2009) used (in addition to traditional tests) well-designed measures in which students had to demonstrate deep understandings of scientific reasoning, and they also failed to find positive effects. The only study of a science inquiry kit that did show positive effects was a very small evaluation of FOSS by Leach (1992). The weighted overall mean effect size across the four studies of science kit programs was only +0.02.

In contrast, several equally inquiry-oriented professional development programs that did not provide kits did show positive science achievement outcomes in rigorous evaluations. These studies provided extensive professional development in effective science teaching, emphasizing conceptual challenge (Mant et al., 2007), cooperative learning (Baines et al., 2007), science-reading integration (Romance & Vitale, 1992, 2011), teaching scientific vocabulary (Rosebrock, 2007), and use of an inquiry learning cycle (Ebrahim, 2004). All eight of these studies found significant positive effects of inquiry-oriented professional development on conventional measures of science achievement, with a weighted mean effect size of +0.30.

The five qualifying studies of technology applications in elementary science all show significant promise. Three approaches had qualifying evaluations: *BrainPOP*, *The Voyage of the Mimi*, and use of web-based laboratory exercises. These applications are all characterized by the use of video or computer graphics to illustrate scientific processes, active inquiry using technology tools, integration of technology, teaching, and group work among students, and efforts to make science content motivating and relevant to students. These science applications are very different from the computer-assisted instruction applications that have dominated uses of technology in elementary mathematics (Slavin & Lake, 2008). Computer-assisted instruction (CAI) in math has emphasized having students work on problems at their appropriate level of need, with feedback on the correctness of their answers, while the science applications with evaluations that met the standards of this review focused more on using technology to enhance classroom teaching and laboratory work.

While the technology applications had the highest weighted mean effect size among the three categories of elementary science approaches (ES=+0.37), it is important to take these findings as promising rather than proven. All of the studies used matching rather than random assignment. Except for the two *BrainPOP* evaluations, the sample sizes are small, and small studies tend to have larger effect sizes than do ones with large samples (Slavin & Smith, 2009). Yet these preliminary findings argue for further development and large-scale evaluations of modern approaches that integrate video and computer technologies with inquiry-oriented teaching and cooperative learning.

Although the limited number of qualifying studies makes explanations of these divergent outcomes tentative at best, it is nevertheless interesting to speculate about their meaning. First, how could the provision of science kits carefully designed to facilitate hands-on inquiry have little benefit for student learning, while other inquiry-oriented professional development approaches did have positive effects? One possible answer may lie in the nature of practical science teaching in elementary schools. In reality, time and resource limitations for elementary science teachers make it difficult to cover the entire science curriculum. In recent years, as high-stakes accountability has focused increasingly on reading and math rather than science, this problem may have become more serious. Elementary teachers who spend a great deal of time on laboratory exercises may be taking time away from coverage of the rest of the science curriculum, especially objectives not covered by the kits. Further, professional development targeted toward helping teachers use kits may not help them enhance their effectiveness on the science units taught without kits.

In contrast, the programs that focus primarily on improving daily instruction on all objectives, not just those that are the focus of provided science materials, may help teachers teach the entire range of science objectives more effectively. That is, a teacher who learns to make effective, daily use of cooperative learning, or conceptually challenging content, or science-reading integration, can take advantage of these new skills every day, for every objective.

If this explanation turns out to be correct, it suggests that elementary science programs might enhance their effectiveness on broadly focused measures of science learning by providing teachers with professional development on methods for teaching all objectives. That is, the findings of the qualifying studies do not call into question the value of inquiry itself or of hands-on laboratory activities, which have long been accepted by the profession as the core of any modern science curriculum (see, for example, Minner, Levy, & Century, 2010; Shymansky, Hedges, & Woodworth, 1990; Bennett, Lubben, & Hogarth, 2006; Anderson, 2002). Yet few if any elementary science teachers use hands-on inquiry activities every day to cover all of the curricular expectations in today's state and national standards. In order to make a substantial difference on broad measures of science learning, teachers may need effective pedagogical strategies for all objectives and all teaching modes.

Far more research and development are needed to identify effective and replicable approaches to improving science achievement outcomes for elementary schools. Science education needs to move beyond brief and artificial pilot tests of exciting new methods and technologies to put them to the test in real schools over extended time periods with valid and comprehensive measures of what students should know and be able to do in science. Science education researchers need to use the tools of science to evaluate and progressively improve the programs and practices needed to help elementary teachers build a scientifically literate society.

References

Anderson, R. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13* (1), 1-12.

Arlin, M., & Webster, J. (1983). Time costs of mastery learning. *Journal of Educational Psychology, 75*(2), 187-195.

Baines, E., Blatchford, P., & Chowne, A. (2007). Improving the effectiveness of collaborative group work in primary schools: Effects on science attainment. *British Educational Research Journal, 33* (5), 663-680.

Bank Street (1984). *The Voyage of the Mimi: Overview guide.* New York: Bank Street College of Education.

Barak, M., Ashkar, T., & Dori, Y. (2011). Learning science via animated movies: Its effect on students' thinking and motivation. *Computers & Education, 56*, 839-846.

Bennett, J., Lubben, F., & Hogarth, S. (2006). Bringing science to life: A synthesis of the research evidence of the effects of context-based and STS approaches to science teaching. *Science Education, 91* (3), 347-370.

Bennett J, Lubben F, Hogarth S, Campbell B (2004) A systematic review of the use of small-group discussions in science teaching with students aged 11-18, and their effects on students' understanding in science or attitude to science. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Borman, G., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness, 1*, 237-264.

Borman, K., Boydston, T., Lee, R., Lanehart, R., & Cotner, B. (2009, March). *Improving elementary science instruction and student achievement: The impact of a professional development program.* Paper presented at the annual meeting of the Society for Research on Educational Effectiveness.

Cobern, W., & Loving, C. (2002). Investigation of pre-service elementary teachers' thinking about science. *Journal of Research in Science Teaching, 39,* 1016-1031.

Cooper, H. (1998). *Synthesizing research* 3rd Ed.). Thousand Oaks, CA: Sage.

Duschl, R.A., Schweingruber, H.A., & Shouse, A.W. (2007). *Taking science to school: Learning and teaching science in grades K-8.* Washington, DC: National Academies Press.

Ebrahim, A. (2004). *The effects of traditional learning and a learning cycle inquiry learning strategy on students' science achievement and attitudes toward elementary science.* (Unpublished doctoral dissertation). Ohio University, Ohio.

Fortus, D. (2008). Science. In T. Good (Ed.), *21st century education: A reference handbook* (Vol. 1, pp. 352-359). Los Angeles: Sage.

Gamoran, A., Borman, G.D., Bowdon, J., Shewakramani, V., & Kelly, K.A. (2012, April). *Implementing district-driven instructional reform: Overcoming barriers to change in a complex urban environment*. Paper presented at the annual meetings of the American Educational Research Association, Vancouver, BC.

Gustafson, B., MacDonald, D., & d'Entremont, Y. (2007). *Elementary science literature review*. Edmonton, Alberta: Alberta Education.

Harlen, W., & Qualter, A. (2008). *The teaching of science in primary schools*. London: Fulton.

Heller, J.I., Daehler, K.R., Wong, N., Shinohara, M., & Miratrix, L.W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching, 49* (3), 233-302.

Hipkins, R., Bolstad, R., Baker, R., Jones, A., Barker, M, Bell, B…..Haigh, M. (2002). *Curriculum, learning, and effective pedagogy: A literature review in science education.* New Zealand Ministry of Education.

Jang, C. (2010). The impact on incorporating collaborative concept mapping with coteaching techniques in elementary science classes. *School Science and Mathematics, 110* (2), 86-97.

Johnson, R., Johnson, D., Scott, L., & Ramolae, B. (1985). Effects of single-sex and mixed-sex cooperative interaction on science achievement and attitudes and cross-handicap and cross-sex relationships. *Journal of Research in Science Teaching, 22* (3), 207-220.

Kilpatrick, J., & Quinn, H. (2009). *Science and mathematics education: Education policy white paper*. Washington, DC: National Academy of Education.

Lazarowitz, R., & Hertz-Lazarowitz, R. (1998). Cooperative learning in the science curriculum. In B. Fraser & K. Tobin (Eds.) *International Handbook of Science Education*. Dordrecht, the Netherlands: Kluwer.

Leach, L. (1992). *Full-Option Science System: Effects on science attitudes and achievement of female fifth-grade students*. (Unpublished doctoral dissertation). Texas Tech University, Texas.

Lipsey, M.W., & Wilson, D.B. (2001). *Practical meta-analysis.* Thousand Oakes, CA: Sage.

Mant, J., Wilson, H., & Coates, D. (2007). The effect of increasing conceptual challenge in primary science lessons on pupils' achievement and engagement. *International Journal of Science Education, 29* (14), 1707-1719.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474-496.

National Research Council (1996). *National Science Education Standards.* Washington, DC: National Academies Press.

National Research Council (2000). *Inquiry and the National Science Education Standards: A guide for teaching and learning.* Washington, DC: National Academies Press.

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

Pell, A., & Jarvis, T. (2003). Developing attitude to science education scales for use with primary teachers. *International Journal of Science Education, 25* (10), 1273-1295.

Pine, J., Aschbacher, P., Roth, E., Jones, M., McPhee, C., Martin, C., Phelps, S., Kyle, T., & Foley, B. (2006). Fifth graders' science inquiry abilities: A comparative study of students in hands-on and textbook curricula. *Journal of Research in Science Teaching, 43* (5), 467-484.

Romance, N., & Vitale, M. (1992). A curriculum strategy that expands time for in-depth elementary science instruction by using science-based reading strategies: Effects of a year-long study in grade four. *Journal of Research in Science Teaching, 29* (6), 545-554.

Romance, N., & Vitale, M. (2011, March). *An interdisciplinary model for accelerating student achievement in science and reading comprehension across grades 3-8: Implications for research and practice*. Paper presented at the annual meeting of the Society for Resarch in Educational Effectiveness, Washington, DC.

Rosebrock, M. (2007). *The effect of systematic vocabulary instruction on the science achievement of fifth-grade science students*. (Unpublished doctoral dissertation). University of Houston, Texas.

Rothman, A. (2000). *The impact of computer-based versus "traditional" textbook science instruction on selected student learning outcomes*. (Unpublished doctoral dissertation). Temple University, Philadelphia, PA.

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.

Schroeder, C.M., Scott, T.P., Tolson, H., Huang, T.-Y., & Lee, Y.-H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching, 44* (10), 1436-1460.

Rothstein, H.R., Sutton, A.J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention assessment, and adjustments.* Chichester, UK: John Wiley.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

SEG Research (2009). *A study of the effectiveness of BrainPOP*. Retrieved January 10, 2012 from www.brainpop.com/about/research.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton-Mifflin.

Shymansky, J. A., Hedges, L. V. and 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,* Vol. 27, No. 2, pp. 127-144.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37(1), 5-14.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78*(3), 427-515.

Slavin, R. E., Lake, C., & Groff, E. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research, 79*(2), 839-911.

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). Effective reading programs for the elementary grades: A best-evidence synthesis. *Review of Educational Research, 79*(4), 1391-1465.

Slavin, R.E. & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness, 4*(4), 370-380.

Slavin, R.E., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31* (4), 500-506.

Sun, K., Lin, C., & Wang, S. (2009). A 3-D virtual reality model of the sun and the moon for e-learning at elementary schools. *International Journal of Science and Mathematics Education, 8,* 689-710.

Sun, K., Lin, Y., & Yu, C. (2008). A study on learning effect among different learning styles in a web-based lab of science for elementary school students. *Computers & Education, 50,* 1411-1422.

Vitale, M. R., Romance, N. R. & Crawley, F. (2010). *Trends in science education research published in the Journal of Research in Science Teaching: A longitudinal policy perspective.* Presented at the annual meeting of the National Association for Research in Science Teaching, Philadelphia, PA.

Vosnidau, S., Ioannides, C., Dimitrakopoulou, A., & Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learning and Instruction, 11,* 381-419.

What Works Clearinghouse (2009). Beginning reading. http://ies.ed.gov/ncee/wwc/

Young, B., & Lee, S. (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), 471-481.

Table 1: Inquiry-Oriented Programs Without Science Kits

Study	Design	Duration	N	Grade	Sample Characteristics	Posttest	Effect Sizes by Subgroup/Measure	Overall Effect Size
Increasing Conceptual Challenge								
Mant, Wilson, & Coates (2007)	Matched	1 year	32 schools (16E, 16C) 1120 students (560E, 560C)	Year 6 10-11 yrs old	Rural and village schools in Oxfordshire, England, mostly White, middle class	(National) Key Stage 2 science tests		+0.33
SPRinG - Cooperative Learning								
Baines, Blatchford, & Chowne (2007)	Matched	2 years	31 schools (12E, 19C) 51 classes (21E, 40C) 1587 students (560E, 1027C)	Years 4-5 8-10 yrs old	Schools in London, England	Items adapted from standardized tests for Year 6, simplified for younger children		+0.21
Science IDEAS								
Romance & Vitale (2011)	Matched	6 years	24 schools (12E, 12C)	3-8	Large urban district in Florida 40% FL 47%W, 29%AA, 19%H	ITBS Science	+0.29	+0.29
						(ITBS Reading +0.40)		
Romance & Vitale (1992)	Matched	1 year	7 classes (3E, 4C)	4	Large urban district in Florida	MAT Science (ITBS Reading +0.40)	+0.90	+0.90
Collaborative Concept-Mapping with Co-Teaching								
Jang (2010)	Matched	8 weeks	114 students (58E, 56C)	4	Science classes in Taiwan	Schoolwide science test on electricity and rainbows		+0.54
Systematic Vocabulary Instruction								
Rosebrock (2007)	Matched	12 weeks	686 students (401E, 205C) Matched on test scores but not demographics	5	Middle-class suburb of Houston	TAKS Earth and Space Science Subtest		+0.24
TEAMS								
Scott (2005)	Matched	1 year	99 students (66E, 33C)	3	Large, diverse district outside of Houston, TX 54%H, 37%AA, 5%W, 83% FL, 40% LEP	ITBS-Science		+0.29
4-E Learning Cycle								
Ebrahim (2004)	Matched	4 weeks	98 students (49E, 49C) in 4 classes	4	Schools in Kuwait	Experimenter-made test--plants and food		+0.96

Table 2: Inquiry-Oriented Programs With Science Kits

Study	Design	Duration	N	Grade	Sample Characteristics	Posttest	Effect Sizes by Subgroup/Measure		Overall Effect Size
Insights, FOSS, and STC									
Pine, Aschbacher, Roth, Jones, McPhee, Martin, Phelps, Kyle, & Foley (2005)	Matched	1 yr	41 classrooms	5	9 diverse school districts in CA, AZ, and NV 500 students in each group	TIMSS Items	-0.02		+0.05
						Performance tasks	+0.11		
FOSS									
Leach (1992)	Random	14 weeks	5 classes (2E, 3C) 103 students (38E, 65C)	5	Urban district in TX, 49% minority	CTBS science	+0.29		+0.48
						Electricity and magnetism test	+0.67		
System-Wide Change for All Learners and Educators (SCALE)									
G. Borman, Gamoran, & Bowdon (2008); Gamoran et al. (2012)	Cluster Random	2 yrs	71 schools (33E, 38C)	4-5	Los Angeles USD 73%H, 11%W, 8%AA 76% FL, 33%ESL		LAUSD Test	State Test	-0.01
						Life Science	-0.04	-0.01	
						Earth Science	+0.01	+0.01	
Teaching SMART									
K. Borman, Boydston, Lee, Lanehart, & Cotner (2009)	Cluster Random	3 yrs	20 schools	3-5	Pasco County, FL	PASS			+0.04
						Multiple choice	+0.08		
						Performance	0.00		

Table 3: Technology Programs

Study	Design	Duration	N	Grade	Sample Characteristics	Posttest	Effect Sizes by Subgroup/Measure	Overall Effect Size
BrainPOP								
Barak, Ashkar, & Dori (2010)	Matched	1 yr	7 schools (5E, 2C) 1335 students (926E, 409C)	4, 5	Students in Israel	Measure based on Israeli national standards		+0.43
SEG Research (2009)	Matched	1 semester	371 students (186E, 185C)	3, 5	Students in Palm Beach (FL) and New York City	SAT-10 Science	Gr. 3 +0.10 Gr. 5 +0.55	+0.33
Voyage of the Mimi								
Rothman (2000)	Matched	1 yr	108 students in 7 classes (57E, 56C)	5	Students in suburban Philadelphia	MAT-7 Science		+0.25
Web-Based Labs								
Sun, Lin, & Yu (2008)	Matched	8 wks	113 students in 2 schools (56E, 57C)	5	Students in Taiwan	Experimenter-made tests of acids and alkalis, use of microscope		+0.30
Sun, Lin, & Wang (2009)	Matched	4 wks	118 students in 4 classes (63E, 65C)	4	Students in Taiwan	Experimenter-made tests of sun and moon systems		+0.26