Experimental Evaluations of Elementary Science Programs:

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
Queen's University, Belfast

Revision, November, 2013

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.

We would like to thank Daphne Minner, Jeanne Century, Derek Bell, Mary Ratcliffe, Judith Bennett, Michael Karweit, and Gavin Fulmer for their comments on earlier drafts.

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 23 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 7 studies), but inquiry-based programs that emphasized professional development but not kits did show positive outcomes (weighted ES=+0.36 in 10 studies). Technological approaches integrating video and computer resources with teaching and cooperative learning showed positive outcomes in a few small, matched studies (ES=+0.42 in 6 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.

Keywords: science education, elementary schools, research review, experimental evaluations

Experimental Evaluations of Elementary Science Programs: 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 Next Generation Science Standards (www.nextgenscience.org/next-generation-science-standards) 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? There has been a rapid increase in the use of rigorous experimental methods to evaluate educational programs of all kinds, and this is beginning to have a significant impact on science education (see Marx, 2012; Penuel & Fishman, 2012). However, experiments evaluating practical applications of alternative science programs and practices are still 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 in real schools over significant time periods.

There have been several reviews of research over time on various aspects of science education, such as inquiry teaching (Anderson, 2002; Bennett, Lubben, & Hogarth, 2006; Furtak, Seidel, Iverson, & Briggs, 2012; 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, and 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, according to the authors. Furtak et al. (2012) identified 22 studies evaluating inquiry methods published in 1996-2006, but only 3 of these involved grades K-6.

While there have been several reviews of research on various aspects of science teaching, there has not been a comprehensive review of experimental evaluations of alternative approaches to elementary science education. 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). A review of research focusing specifically on elementary science approaches is important for several reasons. Science is very different in elementary schools than in middle or high schools, so findings from studies of secondary science may not apply to elementary science teaching. Elementary science is almost always taught by non-specialists, teachers who are responsible for all other subjects and rarely have university degrees in science (Epstein & Miller, 2011). A recent survey (Trygstad, Smyth, Banilower, & Nelson, 2013) found that only 36% of elementary teachers who teach science took at least one university course in life, earth, and physical science, 21% took only one of these, and 6% took none at all. Most teachers reported being very well prepared in reading (80%) and math (77%), but not science (39%). As a result, innovations in science education may need to support teachers' content knowledge and to help them manage limited time and resources. Also, there is little time set aside for science in most elementary schools, and elementary schools rarely have the labs and equipment common in middle and high schools. In the U.S. and the U.K., among other countries, science is not tested as part of state or national accountability until secondary school, so science is often diminished in focus in preference for time and resources devoted to reading and math. In contrast, middle and high schools almost invariably have specialist science educators with regular periods set aside for science, and there is eventually accountability for science learning.

It is important to note ways in which science is different from math and reading, the subjects that have been most often studied using the experimental methods emphasized in this review. First, science covers such a broad range that it is typically taught in time-limited units. For example, a fourth grade teacher might teach a four-week unit on electricity, another on cell functions, and a third on volcanos. These topics do not build on each other, the way skills in reading or math do. Further, there is relatively widespread agreement about the ultimate goals of elementary reading and math instruction, and accountability measures and state standards clearly define what those goals will be. In contrast, the content of science standards is constantly evolving, and is fiercely contested. The lack of science assessments in most states leaves the ultimate goals of science instruction more open to local variation. These aspects of science are important context for this review, and will be discussed in introducing the review methods and explaining the findings.

Affordances and limitations of quantitative reviews in science education. This review applies to elementary science education a quantitative synthesis of experimental studies of practical applications of alternative science approaches. It is important to note that the quantitative review methods applied in this article are well-suited for some questions but not others, and the review does not pretend to encompass all questions. In particular, the focus of this review is squarely on innovations in elementary science, and not on the objectives of instruction. When it is clear what is to be taught, and the question is what materials, methods, and professional development will best accomplish the desired outcome, this review's methods are arguably appropriate, focusing on differential outcomes of different approaches on objectives that all teachers are trying to help their students attain. In contrast, changes in the objectives of instruction, such as emphasis on particular topics, is guided by different imperatives, such as scientific advances and philosophical debates about the purpose of science education, that are less amenable to experimental evaluations.

Science standards develop because scientists, science educators, economists, policy makers, and the general public come to believe that a given topic deserves more attention, or teaching of a given topic at a particular time is believed to facilitate further learning (Wilson, 2009). For example, the recently released Next Generation Science Standards (www.nextgenscience.org/next-generation-science-standards) emphasize the interconnected nature of science, deeper understanding of content, and a greater focus on engineering and technology. They encourage more focus on topics such as climate change and evolution. These standards are statements of a set of understandings and values of what science education should achieve in the modern world. If a student learns more about climate change, that is of value in itself, and teachers teaching about climate change need not be compared to those who do not teach about climate change. What meaningful measure could assess the difference? A test of climate change begs the question, while any other test misses the essence of what the new standards are intended to accomplish.

This distinction between the study of teaching methods and of objectives matters in that many studies in science education evaluate objectives that are different from those emphasized in schools today. Such studies may, for example, provide students with innovative instruction on science topics that students might never otherwise see. They might then give pre- and posttests to note gains made by the students who received the novel content and then compare gains on such measures to those made in a control group. However, in such a study the experimental-control comparison is not a meaningful evaluation, because it is obvious the students taught something that others are not taught at all will learn more of that material. The value of the novel curriculum, in this case, has to be demonstrated in some other way, perhaps using observations, judgments by experts, or international benchmarking. These research methods may be appropriate and rigorous within their own genres.

Even where the objectives are common to all classrooms, having outcome measures that are fair to intervention and control groups is a key methodological consideration. This review excludes studies in which experimenters made their own outcome measures closely aligned with their treatments and then did not ensure that students in the control group were exposed to the content or skills measured on their purpose-built assessments. 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 was more valuable than that learned by the control group, but the experiment does not provide evidence that this particular content is better than traditional content.

Another recent example of the problem of treatment-inherent measures is a study by Heller, Daehler, Wong, Shinutara, and Miratrix (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 been teaching electric circuits at all 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 such comparisons may simply reflect the fact that experimental teachers were teaching about electric circuits during the experimental period and control teachers were not doing so.

A study reported by Slavin and Madden (2011), focusing on math and reading studies reviewed in the U.S. Department of Education's What Works Clearinghouse (WWC), found that measures that are "inherent" to the treatment (covering content not taught in the control group) 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 studies of science education, experimenter-made measures inherent to the content taught only or principally in the experimental condition are often the only measures reported. These measures are often justified by their authors on the basis that the material taught and measured is what students *should* have been taught. This may well be the case, but an experimental study of this kind provides no evidence one way or the other on the value of the experimental treatment.

While recognizing the value of other research methods in science education and, in particular, the importance of arguments for innovations in science standards, the present review focuses exclusively on experiments in which experimental and control groups are equally focused on achieving particular objectives so that they can be fairly compared on common measures. A major limitation of this focus is that most studies that use common measures in experimental and control groups use standardized tests, which many science educators reject as being overly focused on facts rather than inquiry or scientific processes (see, for example, Furtak et al., 2010). Yet there are exceptions, in which more inquiry-oriented measures have been used, and even when this is not the case one can argue that it is of interest to know about science programs capable of improving student outcomes on traditional measures, even as we acknowledge that better measures and better curricula tied to those measures may be desirable. Yet there are exceptions, in which more inquiry-oriented measures have been used, and even when this is not the case one can argue that it is of interest to know about science programs capable of improving student outcomes on traditional measures, even as we acknowledge that better measures and better curricula intended to enhance performance on those measures may be desirable.

Review Methods

The review methods for elementary science applied in this paper are similar to those used in math by Slavin and Lake (2008) and Slavin, Lake, and Groff (2009), and in reading by Slavin, Lake, Chambers, Cheung, and Davis (2009). These reviews used an adaptation of a technique called best evidence synthesis (Slavin, 2008), which seeks to apply consistent, well-justified standards to identify unbiased, meaningful information from experimental studies, and pool effect sizes across studies in substantively justified categories. In these respects, best-evidence syntheses are similar to meta-analyses (Cooper, 1998; Lipsey & Wilson, 2001). That is, they apply consistent inclusion standards to screen all studies meeting initial criteria, and then they use effect sizes (experimental minus control group means divided by the standard deviation of the control group) as a summary of outcomes on each measure. They average effect sizes across studies, weighting by sample size, to obtain estimated treatment effects of practical or theoretical interest. However, what is distinctive to best-evidence syntheses is that in addition to numerical summaries, they provide narrative descriptions of key studies, to give the reader a clear idea of the nature of the original studies, substantive and methodological issues they raise, and findings that go beyond those that are the focus of the review. The intention is to enable readers to understand the reviewers' decisions and to gain insight into the research beyond what meta-analyses ordinarily provide. Further details and rationales for best-evidence synthesis procedures appear in the following sections.

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 (ERIC, Psych INFO, Dissertation Abstracts) using different combinations of key words (for example, "elementary students" and "science achievement") and the years 1980-2012. 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 and Wilson (2001) 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.

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 (comparable standards were used for non-U.S. studies).
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). For example, studies of new technologies that provided graduate students to help all teachers with the technology every day were not included.
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, 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 or posttests with clear indications of ceiling or floor effects were excluded. The criterion for a floor effect was that a measure's standard deviation was larger than the difference between the mean and the lowest possible score. A ceiling effect was defined as existing when a measure's

standard deviation was larger than the difference between the highest possible score and the mean.

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 previously.
8. A minimum study duration of 4 weeks was required. This is much shorter than the 12-week minimum used in the Slavin and 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.

Rationales for Inclusion Criteria

The rationales for most of the inclusion criteria described above are common to most quantitative syntheses (Lipsey and Wilson, 2001). However, there are a few that are worthy of further discussion.

Treatment-inherent measures. The exclusion of studies that used experimenter-made measures of content taught in the experimental group but not the control group was discussed at some length in the introduction. This issue is very important in science education, because many experimental studies in this field use treatment-inherent measures, using the rationale that they cannot find measures they feel to be appropriate to their purposes. Yet as noted earlier, studies find that effect sizes from treatment-inherent measures are far larger than those from measures of content taught equally in all groups (Slavin and Madden, 2011).

Having a measure made by the experimenter did not, however, automatically disqualify a study from the review. In many studies, experimenters made or adapted tests, but they then ensured that both the experimental and the control group taught the same content. For example, Ebrahim (2010) had four teachers each teach two classes on earth, soil, and agriculture for 6 weeks. In each pair of classes, one was taught using cooperative learning and one was taught using conventional methods, but each day the lesson content was otherwise identical, making the experimenter-made test appropriate for both groups.

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 just whether it is proximal or distal.

Study duration. 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, discussed earlier, 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 for practical applications there is little value 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 just a few weeks, we reduced our duration requirement to four weeks, but brief experiments should still be interpreted with caution.

A study by Baines, Blatchford, and 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 article, 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.

Extraordinary resources and artificial conditions. 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, and 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.

Pooling effect sizes. After studies were selected for inclusion, effect sizes were computed for qualifying measures. These were the difference between experimental and control posttest means, adjusted for pretests and other covariates, divided by the control group standard deviation (or the pooled SD if the control group SD was not presented). A mean effect size for the study was then computed, and this mean was pooled, or averaged, with other study means to obtain average effect sizes for various categories of treatments.

In the pooling, studies were weighted by their sample sizes. The reason for this is that research has shown that studies with small sample sizes report much larger effect sizes than larger studies (Rothstein, Sutton & Borenstein, 2005; Slavin and Smith, 2009). Weighting was used to keep small studies from having too much impact on averages. As a consequence, category means are largely determined by the large studies involving many schools, teachers, and students, which are therefore more likely to represent what outcomes might be in large-scale, practical applications of the various programs.

Results

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 332 published and unpublished articles that purported to evaluate science approaches in elementary schools since 1980, only 23 studies met the review standards. (For a table listing studies that did not qualify for the review, and the main reasons they were not included, please contact the first author). 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 elementary science studies that did meet the inclusion criteria provide useful information on several approaches to improving outcomes in science teaching. Seventeen of the qualifying studies focused on *inquiry-oriented instructional processes* for teachers, including approaches such as cooperative learning, integrating science and reading, and use of science kits. 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 (7 studies), and those that provided professional development without kits (10 studies). 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 include 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. It also includes programs such as Scott Foresman Science, which combines kits with leveled readers focusing on science inquiry.

The theory of action underlying the inquiry-oriented programs without science kits, such as cooperative learning and science-reading integration, emphasizes teaching teachers generic strategies they can use every day to make science teaching engaging, comprehensible, and conceptually challenging.

Another category of approaches to improving science instruction emphasizes the use of technological applications to enhance student outcomes. This category includes six studies of 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 ten qualifying studies of interventions in this category. Overall, the sample size-weighted effect size for inquiry-oriented programs that do not use science kits was +0.36.

==========
TABLE 1 HERE
==========

Increasing Conceptual Challenge. Mant, Wilson, and 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.

Cooperative Learning. Two studies evaluated forms of cooperative learning. One of these, by Baines, Blatchford, and 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 second study of cooperative learning was carried out by Ebrahim (2010) in two schools for girls in Kuwait. The cooperative learning intervention was not clearly specified, but it involved organizing students into mixed-ability groups of 4-5. The team method apparently emphasized positive interdependence and individual accountability.

Eight intact fifth grade classes were taught by 4 female teachers (n's = 86E, 78C). Each teacher taught one class randomly assigned to use cooperative learning and one to use teacher-centered instruction during a 6-week unit on earth, soil, and agriculture. Because analysis was at the student level, this was considered a randomized quasi-experiment. Teachers taught the same content in each of their classes.

Students were pre- and posttested on an experimenter-made test on the content taught equally in both types of classes. Controlling for pretests, students in the cooperative learning classes learned significantly more than controls (ES = +0.27, $p<.03$).

Science IDEAS. 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 (2001, 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. Teachers in *Science IDEAS* receive extensive professional development and coaching to help them build comprehension strategies for science and to build

science concepts. Students are taught to link together observed events, to make predictions or manipulate conditions to produce outcomes, and to make meaningful interpretations of events. The science approach emphasizes hands-on activities, concept mapping, and journal writing. In particular, students are 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 visit teachers to monitor fidelity of implementation.

The first large study of *Science IDEAS* was reported by Romance & Vitale (2001). This study involved 15 schools in a diverse district in Florida. A total of 227 students in grades 4-5 were in schools implementing *Science IDEAS,* and 166 were in matched control schools. On MAT Science tests, controlling for ITBS Reading scores from the previous year, students in the experimental classes scored substantially higher (ES = +0.66, p< .01). They also scored better on ITBS-Reading posttests (ES = +0.11, p<.01).

Romance and 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 in the control group had fewer opportunities to use hands-on activities or to pursue science topics in depth.

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.

Controlling for prior-year ITBS reading scores, 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.

Science-Literacy Integration. Another approach to integrating science and literacy in the upper elementary grades was described by Cervetti, Barber, Dorph, Pearson, and Goldschmidt (2012). In their form of science-literacy integration, built around a 40-session (8 week) unit on light, four major investigations were carried out. Each 10-session investigation included 4 days of hands-on activities, 2 of reading, 2 of writing, and 2 of discourse. During the reading sessions, students were taught specific study strategies that they applied in partnership and then in the whole class.

A study of the integrated light unit was carried out with 94 fourth grade classes in a southern state. Teachers were randomly assigned to treatment or control conditions. The groups

were well matched on percent free and reduced lunches (58% E, 53% C), percent African American (36% E, 39% C), Hispanic (6% E, 7% C), and White (53% E, 49% C).

Control teachers were asked to present the content of their state science standards, using their usual methods and materials. One of the four topics taught in the treatment group, Light as Energy, was not taught in the control group, and the control group did not teach a segment on Light and Color, so the assessments focused only on the material covered equally in both groups, on characteristics of light and interactions of light.

Student learning was assessed using an experimenter-made test designed to align with state standards and to fairly assess the content taught in experimental and control groups. Science understanding, Science Writing, Science Vocabulary, and Reading Comprehension were assessed (only the first three relate to the present review). Students were pre- and posttested. On Understanding Science, the posttest adjusted for pretest differences had an effect size of +0.65 ($p< .001$). On Science Vocabulary, the effect size was +0.22 ($p<.001$), and on Science Writing, ES = +0.40, $p<.001$. On reading comprehension, however, there were no differences (ES = +0.09, n.s.).

One concern in this study is whether the results might be due to the treatment simply encouraging teachers to teach more science during the 8 week period, despite the investigators attempts to equalize the focus on science. In fact, teacher surveys indicated that experimental teachers taught science 3.66 hours per week while control teachers allocated 3.03 (ES = +0.53, $p<.05$). They may have particularly spent more time on light. A year-long or at least semester-long study in which teachers in experimental as well as control teachers teach many objectives would be needed to rule out this alternative explanation of the findings.

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. A method for teaching science vocabulary was evaluated in a middle-class suburb of Houston by Rosebrock (2007). 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. A similar proportion of students was economically disadvantaged in both schools (around 16%). 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).

Instructional Process Programs With Science Kits

Instructional process programs that provide teachers with specific materials and instruction 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 seven qualifying studies of instructional process approaches that provide specific student inquiry activities and materials. The sample size-weighted mean effect size for these studies was +0.02, or effectively zero.

==========

TABLE 2 HERE

==========

Insights, FOSS, and STC. Pine et al. (2006) carried out a 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. None of these topics were directly taught in the kits. 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.

AMSTI. The Alabama Math, Science and Technology Initiative (*AMSTI*) is an ambitious, state-wide approach intended to improve performance in upper-elementary and middle schools across Alabama. After developing and beginning dissemination of the program since 2002, the state contracted with a third-party evaluator to do a cluster randomized experiment to evaluate the program starting in 2006-2007 (Newman, Finney, Bell, Turner, Jackiw, Zacamy, and Gould, 2012).

The *AMSTI* intervention involves providing teachers with extensive materials and supplies in kits to enable them to make extensive use of hands-on activities throughout the year. Within regions, kits are rotated among schools every 3-4 months. The kits include equipment such as thermometers, digital cameras, and test kits. Teachers received ten days of inservice (5 in science, 5 in math) during the summer before implementation. During the implementation year, faculty from regional universities visited participating schools to provide encouragement and advice.

The evaluation involved students in grades 4-8 in 82 schools throughout Alabama that had applied to participate. Schools were matched on prior performance and demographic factors, and then one school in each pair was randomly assigned to the experimental group and one to control, yielding 41 in each group. Science was assessed only in grades 5 and 7. Attrition among teachers and students was small and equal across groups. However, the report combines results for grades 5 and 7, so we report them together. Two experimental and one control school

dropped out, leaving 39E, 40C, and there were 192 teachers (102E, 90C) and 7,628 students (4082E, 3446C) in the analytic sample.

The posttest was the routinely administered SAT-10 Science test. SAT-10 Reading was used as a pretest and covariate. The analyses used HLM. The adjusted effect size was +0.05, which was not statistically significant. Results for mathematics and reading were similar.

Despite the disappointing learning outcomes, there was strong evidence that teachers in the *AMSTI* treatment reported more use of active learning instruction in science than did controls (ES = +0.32, p<.002).

SCALE. G. Borman, Gamoran, and 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, and 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 on the performance measures (ES= .00, n.s.).

Scott Foresman Science. *Scott Foresman Science*, published by Pearson, is a year-long curriculum intended to be used every day in grades 3-5. It includes kits focusing on science inquiry, as well as leveled readers. During a half-day inservice, teachers learn to use a strategy emphasizing a progression from "directed inquiry" to "guided inquiry" to "full inquiry." Experiments using materials from the kits are used at all of these stages, but particularly in "full inquiry," where students have the opportunity to work in small groups to set up their own experiments. However, much of class time is spent on the leveled readers, which emphasizes inquiry but does not use the kits.

A third-party evaluator was engaged to evaluate Scott Foresman Science (Miller, Jaciw, and Ma, 2007). The study involved five districts around the U.S. Within each district, teachers of grades 3-5 were randomly assigned to use Scott Foresman Science or to continue with their existing science approaches. The study took place over a full school year. Students were pre- and posttested on the NWEA Science Concepts and Processes and Reading Achievement scales. Both pretests were used as covariates in an HLM analysis. The study involved 36 schools (18E, 18C), 92 teachers (46E, 46C), 113 classrooms (56E, 57C), and 2079 students (1059E, 1020C), who were well matched on pretest and demographic factors. Two of the districts had significant numbers of English learners, but most students were White and spoke English as their first language.

The outcomes indicated no differences on science posttests, controlling for pretests (ES= -0.02, n.s.). There was a small positive effect on reading, but it was also not significant (ES = +0.05, n.s.). Analyzed separately, none of the five districts showed a positive effect.

Project Clarion. *Project Clarion* is a science program for grades K-3 that uses prepared science units from the Integrated Curriculum Model, or ICM (VanTassel-Baska, 1986). Each unit includes an inquiry based on a concept of change or systems. Students take on a role as a scientist, learning the scientific process in order to answer a question or solve a real-world problem.

In a study in six Virginia Title I schools (Kim, VanTassel-Baska, Bracken, Feng, Stambaugh, & Bland, 2012), teachers were randomly assigned to participate in Project Clarion or to serve as a control group. The study took place over 3 years, but since children frequently transitioned between experimental and control conditions, only the first year data could be used. Students in grades K-3 were pre- and posttested on the MAT-8 science test. Adjusting for pretest differences, the posttest effect size was near zero (ES= -0.01).

Technnological Applications

Despite substantial interest in technology applications throughout the science education community and many small trials of exciting innovations, 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 and 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 six studies of technology-focused programs that met the standards of the present review. The weighted mean effect size for these studies was +0.42.

==========

TABLE 3 HERE

==========

BrainPOP. In an Israeli study, Barak, Ashkar, and 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.

Waterford Early Math and Science Program (WEMS). *The Waterford Early Math and Science Program* is an educational software program that provides self-paced computer-assisted instruction in grades K-2. In a study with kindergarten students, Powers & Price-Johnson (2007) evaluated WEMS in five majority-Hispanic schools in Tucson, Arizona. Within the schools, intact classrooms were randomly assigned to use WEMS (n=13 classrooms, 199 students) or to serve as a control group (n=9 classrooms, 139 students). In the WEMS classes, teachers were asked to give students at least four 22-minute sessions each week, split between math and science content. In fact, many students received less than the expected dosage. The researchers established a minimum of 1100 minutes of use as the expectation over the course of the year, and 26% of children did not receive this much exposure.

Students were pre- and posttested on kindergarten forms of the SAT10 Environment test. Adjusting for pretest differences, the WEMS students gained more than controls by an effect size of +0.70.

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 in 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, and 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. 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, and 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 in two treatment and 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, $p<.02$).

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 332 identified studies purporting to evaluate science approaches in elementary schools, only 23 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 of any particular type of program, 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.

It is important to set these studies against the backdrop of contested and changing views about the nature and purpose of science education. Changes in state and national science standards are driven both by advances in scientific knowledge and changing perceptions of what students should be taught. Curricula developed in the 1950s and 1960s when the priority was educating future scientists have been replaced by those emphasising the production of scientifically literate citizens for the twenty-first century (Atkin & Black, 2007; Osborne & Dillon, 2008). Over the same period, there has been a move toward more student-centered, hands-on, dialogic teaching methods (Treagust, 2007). Yet as standards evolve, it is still important to know what programs and teaching methods are most effective in helping students meet whatever standards are currently prevalent.

A surprising finding from the largest and best-designed of the studies synthesized in the present review 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, Project Clarion,* and *Teaching SMART* programs, none of which showed positive achievement impacts. Introduced in the 1990s, these hands-on, kit-based curricula were designed to be easier for teachers to use than the inquiry-based curricula that preceded them. Research has shown that elementary teachers using kits present lessons that are more accurate in content than do those not using kits (Nowicki, Sullivan-Watts, Shim, Young, & Pockalny, 2012). 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 these measures also failed to register 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 six studies of science kit programs was only +0.02.

Previous descriptive research has supported the observation that when teachers are given science kits, their focus can be on implementing the materials rather than on building deeper understandings among students. For example, a large evaluation of the Local Systemic Change Through Teacher Enhancement Program in 61 sites across the U.S. noted that "LSCs have had difficulty in moving teachers beyond 'surface changes'—simply implementing new materials—to the larger task of teaching for understanding" (Boyd, Banilower, Pasley, & Weiss, 2003, p. 64).

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; Ebrahim,

2010), science-reading integration (Cevetti et al., 2012; Romance & Vitale, 1992, 2001), teaching scientific vocabulary (Rosebrock, 2007), and use of an inquiry learning cycle (Ebrahim, 2004). All ten 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 +0.36.

The six qualifying studies of technology applications in elementary science all show significant promise. Four approaches had qualifying evaluations: *Waterford Early Math and Science (WEMS), BrainPOP*, *The Voyage of the Mimi*, and use of web-based laboratory exercises. WEMS is a traditional computer-assisted instruction approach, applied in this case only to kindergartners, but the other three models 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 most of 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=+0.42), it is important to take these findings as promising rather than proven. Most of the studies used matching rather than random assignment, and matched studies leave open the possibility of selection bias (schools or teachers using the programs might have been better or more reform-oriented teachers). 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, for example those 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, 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 so little benefit for student learning, while other inquiry-oriented professional development approaches did have positive effects? One possible answer lies with the nature of the kits themselves, which have been criticized for failing to adequately facilitate conceptual understanding (Boyd et al., 2003). An alternative interpretation relates to 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. Elementary science teachers need to develop pedagogical content knowledge, which means knowing how to make science content meaningful, useful, and engaging (Duschl et al., 2007; Cobern & Loving, 2002; Zembal-Saul, Starr, & Krajcik, 2002). Previous work on cooperative learning in science has demonstrated that it is the interactions established through cooperative learning that best predict positive outcomes (Howe, Tolmie, Thurston, et al., 2002; Thurston,Topping, Christie, et al., 2010).

Many of the science teaching approaches found to be effective in the studies meeting the inclusion criteria resemble methods that have been found to have positive effects in other subjects and in a broader range of science studies. This is particularly true of cooperative learning, which has been frequently found to work at all levels of science education (Bennett et al., 2004; Lazarowitz and Hertz-Lazarowitz, 1998) and in a wide variety of other subjects (Rohrbeck et al, 2003; Slavin, 2013; Webb, 2008). Science-reading integration has also been found to be effective in reading studies (e.g., Guthrie et al., 1998, 1999).

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 fact, research has shown that, despite the focus on inquiry-based teaching in science education policy, it has a generally low profile in classroom practice (Weiss, Pasley, Smith, Banilower, & Heck, 2003). 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.

It is important to note the limitations of this review. Its methods focus on rigorous experimental evaluations of teaching methods and technologies intended to improve the learning of elementary science. However, the review's inclusion standards ruled out many studies that might be of interest to some readers. These include very brief studies (less than 4 weeks), studies that use artificial, non-replicable procedures, and studies in which the control group was not studying the content tested on the outcome measures. These exclusions were intended to focus on studies that inform readers about pragmatic science approaches that could readily be used at a significant scale. However, other experimental, correlational, and observational research is also valuable for theory building, description, and tests of concept. It is not possible in a review to do

justice to every type of research done for every type of purpose, but it would be misleading to suggest that research excluded here is of less importance than studies that were included. Excluded studies simply addressed different objectives.

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. It is encouraging that the framework behind the new U.S. science standards recognizes the need for empirical testing of instructional approaches and specifically recommends randomized trials to evaluate ideas and practices used in the development of learning progressions (National Research Council, 2012). Too many curious, creative students leave elementary school with a diminished love for science and deep misconceptions about scientific principles and the nature of science itself (The Royal Society, 2010). 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.

Atkin, J. M., & Black, P. (2007). History of science curriculum reform in the United States and United Kingdom. In Abell, S. K., & Lederman, N. G. (Eds.), *Handbook of Research on Science Education.* (pp. 781-806). Mahwah, NJ: Erlbaum.

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., 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.

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

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.

Boyd, S., Banilower, E., Pasley, J., & Weiss, I. (2003). *Progress and pitfalls: A cross-site look at Local Systemic Change Through Teacher Enhancement.* Chapel Hill, NC: Horizon Research.

Cervetti, G., Barber, J., Dorph, R., Pearson, P. D., & Goldschmidt, P. (2012). The impact of an integrated approach to science and literacy in elementary school classrooms. Journal of Research In Science Teaching. 49 (5), 631-658.

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.

Epstein, D., & Miller, R.T. (2011). Slow off the mark: Elementary school teachers and the crisis in science, technology, engineering and math education. Center for American Progress, Washington.

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.

Guthrie, J. T., Anderson, E., Alao, S., & Rinehart, J. (1999). Influences of Concept-oriented reading instruction on strategy use and conceptual learning from text. Elementary School Journal, 99(4), 343-366.

Guthrie, J. T., VanMeter, P., Hancock, G. R., Alao, S., Anderson, E., & McCann, A. (1998). Does instruction increase strategy use and conceptual learning from text? Journal of Educational Psychology, 90 (2), 261-278.

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.

Howe, C., Tolmie, A.K., Thurston, A., Topping, K.J., Christie, D., Livingston, K., Jessiman, E., & Donaldson, C. (2007). Group work in elementary science: Towards organisational principles for supporting pupil learning. Learning and Instruction, 17, 549-563.

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.

Kim, K.H., VanTassel-Baska, J., Bracken, B.A., Feng., A., Stambaugh, T., & Bland, L. (2012). Project Clarion: Three years of science instruction in Title I schools among k-third grade students. *Research in Science Education, 42,* 813-829.

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.

Marx, R.W. (2012). Large-scale interventions in science education: The road to utopia? Journal of Research on Science Teaching, 49 (3).

Miller, G., Jaciw, A., & Ma, B. (2007). Comparative effectiveness of Scott Foresman Science: A report of randomized experiments in five school districts. Palo Alto, CA: Empirical Education.

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 framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Washington, DC: National Academies Press.

Newman, D., Finney, P.B., Bell, S., Turner, H., Jaciw, A.P., Zacamy, J.L., & Feagans Gould, L. (2012). Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI). (NCEE 2012–4008). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Nowicki, B.L., Sullivan-Watts, B., Shim, M.K., Young, B., & Pockalny, R. (2012). Factors influencing scientific content accuracy in elementary inquiry science lessons. *Research in Science Education, 43,* 1135-1154.

Osborne, J., & Dillon, J. (2008) Science education in Europe: Critical reflections : a report to the Nuffield Foundation. London: The Nuffield Foundation

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.

Penuel, W.R., & Fishman, B.J. (2012). Large-scale science intervention research we can use. Journal of Research in Science Teaching, 49 (3), 281-304.

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.

Powers, S., & Price-Johnson, C. (2007). *Evaluation of the Waterford Early Math and Science Program for Kindergarten: First-year implementation in five urban low-income schools.* Tucson, AZ: Creative Research Associates.

Rohrbeck, C. A., Ginsburg-Block, M. D., Fantuzzo, J. W., & Miller, T. R. (2003). Peer-assisted learning interventions with elementary school students: A meta-analytic review. Journal of Educational Psychology, 94(2), 240-257.

Romance, N. R., & Vitale, M. R. (2001). Implmenting an in-depth expanded science model in elementary schools: Multi-year findings, research issues, and policy implications. International Journal of Science Education, 23, 373-404.

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 Research 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.

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

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.

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. & 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.

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

Slavin, R. E. (2013). Classroom applications of cooperative learning. In S. Graham (Ed.), APA handbook of educational psychology. Washington, DC: American Psychological Association.

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.

The Royal Society (2010). Science and mathematics education, 5-14: A 'state of the nation' report. London: The Royal Society.

Thurston, A., Topping, K.J., Christie, D., Tolmie, A.K., Karagiannidou, E., & Murray, P. (2010). Cooperative learning in science: Follow-up from primary to high school. International Journal of Science Education, 32(4), 501-522.

Treagust, D. F. 2007. "General Instructional Methods and Strategies." In Abell, S. K., & Lederman, N. G. (Eds.), *Handbook of Research on Science Education.* (pp. 373-391). Mahwah, NJ: Erlbaum.

Trygstad, P.J., Smith, P.S., Banilower, E.R., & Nelson, M.M. (2013). The status of elementary science education: Are we ready for the next generation standards? Chapel Hill, NC: Horizon Research, Inc.

VanTassel-Baska, J. (1986). Effective curriculum and instructional models for talented students. *Gifted Child Quarterly, 30,* 164-169.

Vitale, M. R., Romance, N. R. & Crawley, F. (2010). Trends in science education research

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

Webb, N. M. (2008). Learning in small groups. In T. L. Good (Ed.), 21st Century Education: A Reference Handbook. (pp. 203-211). Los Angeles: Sage.

Weiss, I.R., Pasley, J.D., Smith, P.S., Banilower, E.R., & Heck, D.J. (2003). Looking inside the classroom: A study of K-12 mathematics and science education in the U.S. Chapel Hill, NC: Horizon Research.

What Works Clearinghouse (2013). Procedures and standards handbook (Version 3.0). Washington, DC: Author.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46* (6), 716-730.

Zembal-Saul, C., Starr, M.L., & Krajcik, J. (2002). Constructing a framework for elementary science teaching using pedagogical content knowledge. In J. Gess-Newsome & N.G. Lederman (Eds.), Examining pedagogical content knowledge (Vol. 6, pp. 237-256). Amsterdam: Springer.