Best Evidence
Encyclopedia (BEE)
Empowering Educators with Evidence on Proven Programs
www.bestevidence.org

Effective Secondary Science Programs: A Best-Evidence Synthesis

Alan Cheung
The Chinese University of Hong Kong

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

Elizabeth Kim
Johns Hopkins University

Cynthia Lake
Johns Hopkins University

June 11, 2015

Preparation of this review was partially supported by the Spencer Foundation. We thank Derek Bell, Jeanne Century, Joseph Taylor and Pam Hanley for their comments on an earlier draft.

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

Abstract

This article reports a systematic review of research on science programs in grades 6-12. Twenty-one studies met inclusion criteria including use of randomized or matched assignment to conditions, measures that assess content emphasized equally in experimental and control groups, and a duration of at least 12 weeks. Programs fell into four categories. Instructional process programs (ES=+0.24) and technology programs (ES=+0.47) had positive sample-size weighted mean effect sizes, while use of science kits (ES=+0.05) and innovative textbooks (ES=+0.10) had much lower effects. Outcomes support the use of programs with a strong focus on professional development, technology, and support for teaching, rather than materials-focused innovations.

Key words: middle school science, high school science, secondary science programs, best-evidence synthesis, science review

Effective Secondary Science Programs: A Best-Evidence Synthesis

Every student in America’s schools needs a solid understanding of science (NRC, 2012). Science proficiency is obviously necessary for students intending to pursue careers in the expanding part of our economy that deals with technology, health, environment, engineering, and many other fields. Further, science understanding is an increasingly important requirement for an educated citizenry. Even people who do not take jobs in science-related industries benefit from knowing how science works and what is known due to scientific inquiry, in order to participate knowledgeably as voters and consumers and to maintain their families’ health and well-being (Duschl, Schweingruber, & Shouse, 2007; Kilpatrick & Quinn, 2009).

Despite widespread recognition among policy makers, educational leaders, and the nation as a whole of the importance of science, engineering, and technology as drivers of the future of our country and society, the science achievement of America’s students is mediocre at best, in comparison to that of our international peers. On the 2012 PISA tests in science, U.S. 15-year-olds ranked 28th, slightly below the average of the 65 participating countries. Our PISA science scores are well below those of several Asian countries as well as countries such as Canada, Germany, Poland, the U.K., the Netherlands, and Australia.

On the 2011 National Assessment of Educational Progress (NCES, 2012) science assessment, 68% of eighth graders scored below “proficient.” This was two percentage points better than in 2009, but there is still a long way to go. Further, social class and racial gaps remain substantial. While 57% of White students and 59% of Asian/Pacific Islanders scored below proficient, corresponding percentages were 90% for African Americans, 84% for Hispanics, and 80% for American Indians. Among students eligible for free or reduced-price lunches, 84% scored below proficient, compared to 55% of non-eligible eighth graders. In other words, science achievement is a serious problem for all eighth graders, but it is a crisis for poor and minority students. A student who scores below proficient in science is very unlikely to seek or qualify for post-secondary education with a STEM focus or careers in STEM.

In recent years, there has been growing consensus about the goals of science education in elementary and secondary schools. This consensus is captured in particular in the Next Generation Science Standards (www.nextgenscience.org/next-generation-science-standards) and the National Science Education Standards (National Research Council, 1996, 2000, 2012). In particular, science leaders agree that science teaching should emphasize conceptual understanding rather than facts, and that inquiry-oriented teaching should be expanded. Yet there is a wide gap between agreement on standards and agreement on specific teaching methods and programs. Standards can help teachers, textbook authors, and program developers identify topics to emphasize, but what kinds of approaches can teachers use to ensure that students will succeed on the new standards?

In parallel with the recent reforms in standards, there has also been an acceleration in the use of rigorous quantitative methods to evaluate innovative science methods (Marx, 2012; Penuel & Fishman, 2012). These are studies that compare classes using innovative methods to those using traditional or alternative methods. Studies are increasingly using random assignment to conditions, large samples, long durations, and other features that add rigor and reduce bias in experimental studies.

Of course, experimental studies of science teaching innovations have long been done, but most have been brief and artificial experiments, have lacked control groups, have used measures closely aligned to experimental but not control treatments, or have otherwise allowed for the possibility that outcomes will greatly overstate the impacts teachers will actually observe in practical applications. Recent reviews of research on science education by Schroeder, Scott, Tolson, & Lee (2007), Furtak, Seidel, Iverson, & Briggs (2012), and Minner, Levy, & Century (2010) have summarized evidence from studies of elementary and secondary science teaching, especially inquiry methods, but all have acknowledged the substantial diversity in the nature and quality of program evaluations in science education. When different types of programs are confounded with differing evaluation methods, it is difficult to draw well-justified substantive conclusions.

As the number of rigorous experiments evaluating various science approaches increases, it becomes both necessary and possible to use accepted meta-analytic methods to summarize the findings of studies meeting clearly specified criteria. It is possible to learn from any study, but findings from studies that compare the gains of experimental and control students on fair and valid measures are likely to be of particular value in building a science of educational practice, and to offer solid evidence to education leaders and policy makers about the types of interventions most likely to increase student achievement.

Slavin, Lake, Hanley, & Thurston (2014) reviewed rigorous quantitative research on science programs at the elementary level. The elementary review emphasized three types of programs: inquiry-oriented programs without science kits (e.g., *Increasing Conceptual Challenge*, *Science IDEAS*), inquiry-oriented programs with science kits (e.g., *Insights, FOSS*) and technology programs (e.g., *BrainPop, Voyage of the Mimi*). Study inclusion criteria included the use of randomized or matched control groups, an intervention duration of at least 4 weeks, and the use of achievement measures independent of the experimental treatment. Twenty-three studies met these criteria. Programs that used science kits did not show positive outcomes on science achievement measures (weighted ES=+0.02 in 7 studies). However, inquiry-based programs that emphasized professional development but not kits did show significant positive outcomes (weighted ES=+0.36 in 10 studies). The largest effect sizes came from a small number of studies of technology approaches integrating video and computer resources with teaching (weighted ES=+0.42 in 6 studies). The review concluded that science programs focused on improving teachers' classroom instruction, such as cooperative learning and science-reading integration, as well as programs making innovative uses of technology closely integrated with teachers' instruction, are promising avenues for improving science teaching and learning.

Focus of the Current Review

The present review uses procedures similar to those used by Slavin et al. (2014) to review research on science programs for middle and high schools, grades 6-12 (ages 11 to 18). Sixth graders appeared in the earlier review if they were in elementary schools, in the current review if they were in middle schools. As in Slavin et al. (2014), the intention of the present review is to place all types of programs designed to enhance the science achievement of middle and high school students on a common scale, to provide educators with meaningful, unbiased information that they can use to select programs most likely to make a difference for their students' learning. In addition, the review is intended to look broadly for common factors that might underlie effective practices across programs and program types, and to inform an overarching understanding of effective instruction in secondary science.

This synthesis also seeks to identify common characteristics of programs likely to make a difference in student science achievement. It includes all types of approaches to science instruction, grouping them in four categories. *Instructional process* approaches were ones that provided substantial training and coaching to teachers in specific approaches to inquiry-oriented science teaching, such as cooperative learning and metacognitive strategy instruction.

Technology programs are ones that make extensive use of computers to enhance science learning. *Science kits* are programs that provide teachers with inquiry-oriented kits facilitating hands-on experiments, as well as extensive professional development to use the kits. *Textbook programs* provide innovative or standards-based content, but far less PD than instructional process approaches and minimal or no use of technology. Note that these categories differ from those used by Slavin et al. (2014) because the qualifying secondary studies evaluated a somewhat different set of approaches.

Readers may wonder about the role of inquiry teaching in these categories. Because “inquiry” is so widely claimed by innovators, it is not a useful criterion in itself for categorizing science programs. In science education, virtually all innovative approaches proclaim their support for inquiry (Duschl, 2003, 2008;Furtak et al., 2012; Minner at al., 2010; Schroeder et al., 2007). Yet the meaning of inquiry in practice varies substantially. Inquiry usually implies extensive use of open-ended experiments, problem solving, and simulation, but virtually every textbook, no matter how traditional, describes itself as inquiry-based.

Method

The review methods for secondary science applied in this paper are essentially the same as those used by Slavin et al. (2014), except that they required a 12-week study duration rather than a 4-week duration. These were in turn adapted from previous reviews of studies in elementary and secondary mathematics and reading (Slavin & Lake, 2008; Slavin, Lake, & Groff, 2009; Slavin et al., 2009; Slavin et al., 2008). All of 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 (covariate-adjusted experimental control group mean differences divided by the unadjusted standard deviation) 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 programs and studies, and to gain insight into the research beyond that which 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, “secondary science” and “science achievement”) and the years 1990-2015. Results were then narrowed by subject area (for example, “educational software,” “secondary education,” “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 secondary 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, 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 secondary science) and basic methodological characteristics (i.e., they had a well-matched control group and a duration of at least 12 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 the same as those used by Slavin et al. (2014), except that students had to be in middle or high schools, and that the minimum duration was 12 rather than 4 weeks.

1. The studies evaluated programs and practices used in secondary science, and were published in 1990 or later. Studies could have taken place in any country, but the reports had to be available in English.
2. The studies took place in middle and high schools.
3. The studies compared students 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. Random assignment could be at the level of individuals or clusters (e.g., schools or classes). If random assignment was at the cluster level but there were too few clusters for analysis accounting for clustering, this is termed a "randomized quasi-experiment" (Slavin, 2008) and categorized as a matched study. Studies without control groups, such as pre-post comparisons and comparisons to "expected" scores, were excluded.
6. Pretest data had to be provided, and there could be no indications of initial inequality. If science pretests were not available, standardized reading or math tests 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 could also be excluded based on other important differences, such as comparing students using a given program in magnet schools to controls in non-magnet schools.

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.
8. A minimum study duration of 12 weeks was required. This was the same as the duration criterion used in all mathematics and reading reviews by the same authors, but not for the elementary science review, which required a 4-week duration. The shorter duration was used in the elementary review because there were many studies at that level focusing on a single unit of a few weeks' duration.
9. Studies had to have at least two teachers, two schools, and 15 students in each treatment group. This criterion reduced the risk of confounding teacher, class, or school effects with treatment effects.

Results

Study Characteristics

A total of 21 qualifying studies based on over 31,000 students in grades 6-12 met the inclusion criteria. Ten studies were quasi-experiments (including randomized quasi-experiments) and 11 were randomized studies. Findings were reported in 16 published articles and five in unpublished documents such as dissertations and technical reports. Two of the studies were published in the 1990s, 8 in the 2000s, and 11 in the 2010s. In terms of sample size, eight were small-scale studies (*N*<250) and 13 were large-scale studies (*N*≥250). These studies covered a wide range of science subjects, including physics, chemistry, biology, and general science. Eleven of these studies examined the program impacts on middle or junior high school students and ten on senior high school students.

Overall Effects

The overall sample size-weighted mean effect size for the 21 qualifying studies was +0.21. A large *Q* value (Q_B=136.09, *df*=20, p<0.000) suggests that there are substantial variations in outcomes among this collective set of studies. We will present the findings according to substantive and methodological features to model some of these variations in the following sections.

Instructional process programs. Instructional process science approaches are ones that provide teachers with extensive training and/or classroom coaching in specific teaching methods, such as cooperative learning, use of metacognitive strategies, and project-based learning. Methods of this kind that have been evaluated in qualifying studies do not provide generic professional development on science content or refinements on current instructional methods, but instead have well-defined models of what classroom science teaching should look like and use extensive professional development to help teachers adopt and implement the innovative models. Instructional process models vary substantially, so they should not be seen as a consistent strategy. What they share is a theory of action emphasizing improving student science learning by providing extensive professional development designed to change teachers' behaviors, rather than focusing primarily on innovative materials, texts, or technology with limited PD. Most of the programs in this category provide at least five days of in-service, plus on-site follow-up in many cases, making them very different from textbook approaches.

Seven programs in the instructional process category each had one qualifying evaluation. The weighted mean effect size was +0.24. Table 1 summarizes the study characteristics, designs, and findings for this category.

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

TABLE 1 HERE

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

Peer-Mediated vocabulary intervention. In a matched control study, Green (2010) evaluated the effects of providing professional development in a peer-mediated vocabulary strategy on science achievement in four middle schools in a southeastern state. The study involved eight teachers and 675 seventh grade students with and without learning disabilities in 41 classes. Teachers were randomly assigned to either experimental or comparison conditions. In comparison classrooms, teachers often began the class with a warm-up activity such as doing a worksheet and reviewing previously learned materials. After the warm-up activity, teachers presented new units. In the treatment classrooms, students received the same types of instruction except that for two days a week, students used the first 15 minutes to learn new science terms by working with their partner using a pair routine and researcher-developed science vocabulary cards.   On a third day, the teachers gave them a short assessment on these new science terms The findings showed that the treatment group that used the peer-mediated vocabulary intervention outscored the controls on the science standards-based assessment developed by the researcher with an adjusted effect size of +0.24.

IMPROVE. IMPROVE is a cognitive-motivational self-regulation approach in which teachers are given professional development in strategies to help students learn to think themselves through difficult problem solving. It has been successfully evaluated in mathematics in studies in Israel (Mevarech & Kraminski, 1997), and was extended into science education and evaluated by Michalsky (2013) as a means of helping students read science texts. Students work in small groups to learn four self-questioning strategies (comprehension, connection, strategy, and reflection) to ask themselves questions such as “what is the phenomenon all about? What do you already know? What are the similarities and differences between this problem and others you know about? Does a given solution make sense?” Motivational questions are also asked for each of these, such as, “What is your motivation for solving the problem or doing the task?”

Michalsky (2013) evaluated the full IMPROVE model and two variations in comparison to a control group in 10th grade biology classes in Israel. All classes studied the same 12-week unit on micro-organisms using the same book. Eight teachers were randomly assigned to the four conditions (two per condition), but analysis was at the student level, so this was a randomized quasi-experiment. Students were pre- and posttested on items from the PISA science literacy test relating to general high school scientific literacy. Controlling for pretests, students in the full IMPROVE condition (*n*=49) scored substantially higher than those in the control group (*n*=46) (ES=+1.26). Students in a cognitive-only group (*n*=50) had an effect size compared to controls of +0.72, and those in a motivation-only group *(n*=53) had an effect size of +0.64.

BSCS inquiry approach. BSCS developed an inquiry-focused approach to science teaching in grades 9-11, including integrated content from life science, physical science, Earth-space science, and the history and nature of science. The approach combines curriculum materials with extensive professional development, seven days for each teacher every year.

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

Taylor, Getty, Kowalski, Wilson, Carlson, & Van Scotter (2014) carried out a two-year cluster randomized evaluation of the BSCS Inquiry Approach in 18 high schools in Washington State. The sample was ethnically diverse, and 45% qualified for free lunch. Schools were randomly assigned to BSCS (*n*=1509) or control (*n*=1543) conditions in 10th and 11th grade. Data were analyzed using HLM. On Washington State Science assessments, controlling for pretests, students in the BSCS schools scored significantly better than controls, with an effect size of +0.09.

Project-Based Inquiry Science (PBIS). Harris, Penuel, DeBarger, D'Angelo, and Gallagher (2012) conducted a large-scale cluster randomized study to examine the effectiveness of *Project-Based Inquiry Science*, a comprehensive, 3-year middle school science curriculum. The program was designed to promote student engagement with science and engineering practices by using models, constructing explanations, and engaging in argument from evidence. This one-year long randomized controlled trial involved about 2,400 sixth grade students from 42 middle schools in one large highly diverse urban school district. Schools were randomly assigned to treatment or control conditions. Participating teachers in both treatment and control conditions taught the same science topics in physical science and in Earth science. The control condition used the district-adopted textbook, McDougall Littell's Science (2005) for Grade 6. The treatment teachers participated in a series of PBIS curriculum-focused workshops provided by the publisher three times a year. At posttest, the treatment students scored significant higher than the control group on both the physical science (energy) unit and the Earth science unit with effect sizes of +0.21 (p<0.04) and +0.25 (p<0.06), respectively. The combined effect size across the two measures was +0.23.

Prediction/discussion-based learning cycle instruction. Lavoie (1999) carried out a 3-month study to examine the effects of giving teachers professional development in the use of a prediction/discussion-based learning cycle (HPD-LC) in high school biology classes. The study involved a total of five teachers and 250 tenth graders who were of lower to middle socioeconomic class. All five teachers were asked to teach one treatment class (HPD-LC) and one control class. The five treatment and five control classes were matched based on class size and student ability. In addition, the pretest scores also indicated that the two groups were very comparable. The HPD-LC approach consisted of three learning phases. In the first phase, students were encouraged to write out their predictions with explanatory hypotheses. Students were then asked to engage in an interactive debate involving reasons. The final phase required students to "solve problems and answer questions that relate to and extend the concept identified in the previous phases" (p. 1130). After the 3-month intervention, three outcome measures were used to assess their achievement. The effect sizes were +0.42, +0.40, and +0.56 for Processes of Biological Investigation Test, Group Assessment of Logical Thinking, and Concept Understanding Test, respectively. The overall effect size across the three measures was +0.46.

Making Sense of SCIENCE.^{TM} In a large-scale cluster-randomized study, Heller (2012) evaluated the effectiveness of the *Making Sense of SCIENCE*^{TM} professional development course on force and motion. Participants were over 6,500 eighth grade students, 181 teachers from 137 schools in 55 diverse districts in California and Arizona. Teachers were randomly assigned to treatment or control conditions (90E, 91C). Treatment teachers received training from the research team on various effective science teaching strategies, including inquiry-based science instruction, hands-on activities based on natural phenomena, and collaborative, small-group work. After one year of intervention, no significant differences were found between the two conditions on the ATLAS Test of Force and Motion (ES=+0.11) and the California Standards Test-Physical Science (ES=+0.03). The average effect size across the two measures was +0.07.

Integrated, activity-based science curriculum. Turpin (2000) carried out a quasi-experiment to examine the effects of an integrated, activity-based science curriculum on science content achievement. Seventh grade science students (532E, 450C) from seven schools in a mid-size city in north Louisiana were chosen to participate in this one-year long study. The treatment group received instruction which emphasized hands-on science activities that engaged scientific thinking and reasoning. The control group used a traditional science curriculum that used lectures and demonstrations as their primary method of instruction. Science achievement was measured using the science subtest of the ITBS. At the posttest, the treatment group scored significantly higher than the controls on the ITBS with an adjusted effect size of +0.16.

Technology programs. Technology programs, of course, have in common the extensive use of digital devices. Their theories of action emphasize the power of digital media to provide material appropriate to students' needs and to integrate visual and text elements of science concepts. Remarkably, none of the technology approaches identified for this review used the computer-assisted instruction (CAI) strategies that have dominated the use of technology in math and reading for many years, which research has not generally supported (see Cheung & Slavin, 2013 a, b; Slavin, Lake, & Groff, 2009). In science, at the elementary as well as the secondary levels, technology is more often used to illustrate science concepts, simulate real-world processes, and support the teacher's instruction, rather than operating separately.

Only five programs, each with just one evaluation, met the inclusion requirements, but these studies had by far the most positive average effect size among the four categories of programs in the review. The weighted mean was +0.47, even higher than the mean of +0.37

reported for the six qualifying technology studies in elementary science by Slavin et al. (2014). The study characteristics and outcomes are shown in Table 2.

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

TABLE 2 HERE

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

The focus of the technological applications in secondary science mostly included strategies designed to help students visualize science concepts, and to connect to resources beyond the classroom. Four of the five qualifying studies were small (n<250) and all five used matched rather than randomized assignment to conditions, so these findings are far from conclusive.

BrainPOP. A third-party evaluation by SEG Research Group (2009) evaluated *BrainPOP*, a web-based multimedia learning application designed to deliver instructional content using two main modes: visual and auditory. The theory behind BrainPOP is that students learn better when the materials are presented in both words and pictures than in words alone. BrainPOP can be used in both group and one-on-one settings. Participants were 128 eighth graders (59E, 69C) from schools in Palm Beach, Florida and New York City. After controlling for pretest difference, the treatment students outperformed the controls on SAT 10 Science, with an adjusted effect size of +0.41.

Constructing Physics Understanding (CPU). Huffman, Goldberg, and Michlin (2003) evaluated the *Constructing Physics Understanding Project (CPU)*, a program that uses computer-based modular curricular activities, software, and pedagogy to help teachers create constructive learning environment for their students. The study involved 13 teachers (8E, 5C) and 194 students (116E, 78C) from 23 high school physics classes in which force and motion units were taught. At the end of this one-year study, both groups were administered a nationally recognized test of force and motion, the *Force Concept Inventory* (FCI). After adjusting for pretest differences, the effect size was +0.64.

Integrated video media. Harwood and McMahon (1997) reported an evaluation of an integrated video media curriculum enhancement in a first-year high school chemistry course. The study took place in a multi-culturally diverse metropolitan region of the East Coast. Participants were 450 first-year general chemistry students in 18 classrooms. The seven treatment classes were taught micro-unit chemistry topics along with *The World of Chemistry*, a video series produced by the University of Maryland, College Park. The series was designed to "bring abstract, distant worlds of science into close focus and within the personal realm of each individual student" and "enable the teacher to stop the videotape approximately every 5-7 min for a teacher-student question-answer interaction time" (p. 620). Control classes used traditional textbooks. The High School Studies Test: Chemistry, a 40-min standardized norm-referenced test, served as both pretest and posttest. At posttest there were significant and substantial differences on the High School Subject posttest, with a pretest- adjusted effect size of +0.71.

iBooks. iBooks is an Apple application that enables students to download and read textbooks on an iPad. In addition to straight text, iBooks allow students to customize their reading using functions such as highlighting, searching, note taking, and bookmarking. It allows students to zoom into features, touch parts of pictures to get additional explanations, and access videos to add explanations and add context. Pearson partnered with Apple to create a digital biology approach based on the Miller & Levine biology text.

Pearson researchers (Baughman, Ehmann, & Vilcheck, 2013) carried out a matched evaluation of the iBook biology program. Three teachers, two in New Jersey and one in South

Carolina, were asked to designate two similar classes and select one to use iBooks and one to use the print texts they had used previously. Students were pre- and posttested on SAT 10-Science. Adjusting for pretests, students using the iBooks gained significantly more than those using print books (ES=+0.25).

1:1 laptop use Dunleavy and Heinecke (2007) evaluated the impact of 1:1 laptop use on middle school science and math standardized test scores in a struggling urban middle school in a mid-Atlantic state. Students (52E, 111C) were randomly assigned to one of the two conditions. Treatment students received a laptop, an Apple iBook with 128 megabyte hard drives running Operating System 10.2.8. Each laptop was loaded with online mathematics and science textbook access and laptop-based instruction. The control group had access to all the resources allocated to the treatment group in a school computer lab. The intervention took place over a period of two years. A standardized math and science test was used to examine the outcomes. Significant differences were found on the standardized science test favoring the treatment group (ES=+0.24, p<0.03).

Science Kits. Just one study by Newman et al. (2012) evaluated a program in which teachers were given science kits to help students do experiments and other inquiry-oriented activities. Unlike textbook programs, such approaches provide a great deal of professional development to teachers, and unlike professional development programs they provide specific, well-developed classroom materials that engage students in hands-on experiments and explorations. Characteristics and outcomes of this study appear in Table 3.

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

TABLE 3 HERE

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

Alabama Math, Science, and Technology Initiative (AMSTI). In a large-scale cluster randomized experiment, Newman et al. (2012) evaluated the effects of the Alabama Math, Science, and Technology Initiative (AMSTI). The study involved 7,258 grade 5 and 7 students and 780 teachers in 79 schools from five regions of Alabama. AMSTI is a two-year intervention aimed at enhancing the alignment between classroom practices and national and statewide standards, and ultimately improving student achievement. The initiative provided teachers with professional development, access to materials and technology, and in-school support. The focus of professional development was to enhance teachers' ability to use higher levels of hands-on, inquiry-based instruction. Though this was a 2-year study, only first-year results were valid because the control group started to implement the AMSTI program after the first year. For grades 5 and 7, data were not broken down by grade, so Table 3 shows the total sample, recognizing that this combines upper elementary and middle school students. Schools were matched based on demographics and prior achievement level and were then randomly assigned to the treatment or control condition. The effect of AMSTI on SAT 10 Science Achievement after one year was statistically significant but very small (ES=+0.05).

Textbooks. Textbook innovations represent an approach to science education reform emphasizing the content of courses. The theory of action behind textbook approaches assumes that standards-based content or other features of texts will improve student science outcomes. Professional development is invariably provided to teachers to help them use new textbooks, but on the order of two to five hours at most, in contrast to the five days or more typical of professional development approaches. Also, in textbook methods the innovation is in the content rather than the teacher's instructional methods, which is why less professional development is provided. Eight studies of five textbook innovations found very small impacts (weighted mean effect size =+0.10) for innovative textbooks. Characteristic of these studies are presented in Table 4.

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

TABLE 4 HERE

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

Miller & Levine Biology. Eddy and her colleagues carried out three studies (Eddy & Berry, 2005; Eddy & Berry, 2007; Eddy, Ruitman, Sloper, & Hankel, 2010) to evaluate the effectiveness of Miller and Levine's Biology curriculum. The program was designed to promote optimal student learning by using real-world applications, hands-on activities, differentiated instruction, sequenced student assessments, and other inquiry activities.

The first study was a small-scale matched control pilot study (Eddy & Berry, 2005). Participants were 205 9th and 10th graders (92E, 113C) from 4 classes in an ethnically diverse high school in California. The outcome measure was a standards-based, nationally recognized biology assessment. No significant differences were found at posttest between the treatment and the control classes, controlling for pretests.

Eddy and Berry (2007) later conducted a cluster randomized trial (RCT) on the Miller & Levine program. Approximately 1,100 high school students and 16 teachers from five high schools in four states (California, Colorado, Ohio, and New Jersey) were involved in this one-year long study. Teachers were randomly assigned to either the treatment or control condition at each study site. The control teachers used the biology curriculum currently in place in their school. HLM analyses found no statistically significant difference between the two groups (ES=+0.02, n.s.).

Eddy, Ruitman, Sloper, and Hankel (2010) carried out a cluster randomized trial in which teachers were randomly assigned to either the treatment or control conditions at each study site. Twenty-four teachers and almost 2,000 students from six schools across five states participated in this one-year study. After adjusting for pretest differences the treatment group scored non-significantly higher than the control group on SAT-9 (ES=+0.18) and Biology Content Assessment (ES=+0.02) with an average effect size of +0.10.

Harcourt's Holt McDougal Biology Program. To evaluate the efficacy of the Holt McDougal Biology program, Shannon and Grant (2012) conducted a cluster randomized trial in which teachers were randomly assigned to either treatment or control conditions. Holt McDougal Biology is a high school biology program, which uses a combination of textbook, online, and multimedia resources designed to promote student interest in biology. A total of 24 teachers and over 1,400 high school students (majority 9th and 10th graders) participated in this one-year long study. The adjusted effect sizes for SAT 10 Science Achievement and PASS Biology Achievement were +0.06 (n.s.) and +0.12 (n.s), with an average effect size of +0.09.

Prentice Hall Science Biology. Two studies were carried out to examine the effectiveness of *Prentice Hall Science Explorer* (Prentice Hall, 2003; Resendez & Azin, 2006). The program was designed to develop and sharpen students' inquiry abilities such as observing, inferring, and graphing.

The first study was a quasi-experiment conducted by the publisher. In the beginning of the school year, students were tested with the TerraNova CTBS Complete Battery Plus. At the end of the study, students were retested with the same test. Two hundred and twenty-three eighth grade students (108E, 115C) from six schools across four states (CO, NJ, WA, & WI) participated in the study. Though the treatment group scored higher than the control group at posttest (ES=+0.12), the difference was not statistically significant.

The second study was carried out by Resendez and Azin (2006). Seventeen teachers and 1,255 sixth to eighth grade students from four geographically dispersed schools were involved in this one-year cluster randomized study. Teachers were randomly assigned to treatment (*n*=10) and control (*n*=7) conditions. After adjusting for pretest differences, non-significant effect sizes for ITBS and TIMSS were -0.04 and +0.12, respectively, for a mean of +0.04.

Pearson Interactive Science Program. Thirty-five teachers and 1,362 sixth to eighth grade students from nine geographically dispersed schools participated in a year-long cluster randomized study that examined the effects of the Pearson Interactive Science program on students' achievement (Resendez, DuBose, & Azin, 2011). The program was designed to promote higher-order thinking skills and real-world connections. Teachers were randomly assigned to either treatment or control conditions. At posttest, adjusting for pretests, the treatment group had significantly higher scores than the controls on the TerraNova Science Test with an adjusted effect size of +0.14

Houghton Mifflin Harcourt's Science Fusion. Resendez and Azin (2013) carried out a 2-year longitudinal cluster randomized trial to examine the effectiveness of *Harcourt's Science Fusion*, a middle grades science program designed to promote higher-order thinking skills and student engagement. Sixth and seventh grade students (*n*=576) from 27 classes in three schools participated in the study. Classes were randomly assigned to either treatment (*N*=14) or control conditions (*N*=13).

At posttest, the treatment classes scored significantly higher than the controls on ITBS with an effect size of +0.39.

Outcomes by Substantive and Methodological Features

Categories of science programs. The four categories of science programs (instructional process, technology, science kits, and textbooks) differed substantially from one another in outcomes (Q_B = 30.06, *df*=3, *p*<.001)

Grade levels. We examined whether there were any differential impacts at different grade levels. Eleven of the studies involved middle/junior high school students (Grades 6-8) and ten involved senior high school students (Grades 9-12). The effect sizes for middle school and high school were +0.15 and +0.30, respectively. This difference was marginally significant (Q_B=3.39, *df*=1, *p*<0.07).

Experimenter-made measures. In this set of studies, we found significant and substantial differences between studies that used experimenter-made measures (including studies in which experimenters chose items from standardized tests) and those that used whole standardized tests (Q_B=4.98, *df*=1, p<0.03). The effect sizes for experimenter-made measures and standardized tests were +0.45 and +0.16, respectively. Note that this difference was observed even though studies using experimenter-made measures had to provide evidence that experimental and control groups were exposed to the same objectives.

Publication bias. To examine the possible impact of publication bias, we carried out two statistical analyses: Classical fail-safe N test and Orwin's fail-safe test. The results from the classical test indicated that in order to nullify the overall effect size of +0.21, a total of 686 studies with null results would be needed. The Orwin's test also generated similar findings. In order to bring the existing overall mean effect size to a trivial level (essentially zero), the number of null studies would have to be 193. The findings of these two tests provide clear evidence that publication bias could not account for the positive effect size seen across all studies.

We used a mixed-model method to examine whether differences existed between published articles and unpublished reports such as technical reports and dissertations. The mean effect sizes for published articles and unpublished reports were +0.64 and +0.12, respectively. The difference is consistent with those of other meta-analyses (e.g., Cheung & Slavin, 2013a; 2013b; see also Lipsey & Wilson, 2001).

Research design. Effect sizes may also vary according to the nature of study research designs. Previous reviews have indicated that matched studies generally produce much larger effect sizes than randomized studies (Cheung & Slavin, 2012; 2013a, b). For example, when examining the impact of educational technology approaches on reading and mathematics achievement, Cheung and Slavin (2012; 2013) found that the effect sizes were about twice as large in quasi-experiments (including randomized quasi-experiments) than in randomized experiments. We found similar results in secondary science. The mean effect size for the 10 qualifying matched control studies was +0.40, whereas the mean effect size for the 11 randomized studies was +0.10.

Sample size. Another potential source of variation may have to do with sample size (Slavin & Smith, 2009). Previous meta-analyses suggest that small studies usually produce much larger effect sizes than large studies (Cheung & Slavin, 2012; 2013a; 2013b; Liao, 1999). A statistically significant difference was found between large studies and small studies (Q_B=4.45, *df*=1, *p*<0.03). The effect size for the 14 studies with large sample sizes (*N*>=250) was +0.16, and the effect size for the seven studies with small sample sizes was +0.39. The results should not come as a surprise because it is easier to maintain high implementation fidelity in small-scale studies than in large-scale studies. In addition, standardized outcome measures are more likely to be used in large-scale studies, which are often less sensitive to the treatment. Furthermore, small studies with null results are less likely to be published or made available in report form.

Discussion

The findings of the present review correspond remarkably well with those of the Slavin et al. (2014) review of elementary science programs. As in the elementary review, the most positive effect sizes among qualifying studies were associated with innovative uses of technology (weighted mean effect size = +0.47) and, to a much lesser extent, instructional process programs making extensive use of professional development (weighted mean effect size = +0.24). In contrast, the one program emphasizing use of science kits (effect size = +0.05) and those providing alternative textbooks (weighted mean ES = +0.10) had minimal impacts. In the elementary review, technology (weighted ES=+0.37 across five studies) and instructional process programs without science kits (weighted mean ES = +0.36 across 10 studies) had positive effects, while science programs using kits had no effect on learning (weighted mean ES=+0.02 across five studies). No studies of textbooks qualified for the elementary review, although a large study of Scott Foresman Science, categorized as a science kit approach, had an effect size of -0.02 (Miller, Jaciw, & Ma, 2007).

The findings of these reviews are so similar and so striking that it is useful to consider them together. The number of studies meeting the inclusion criteria in each review is small (23 in the elementary review, 21 in secondary), so using the full set of 44 studies helps to see patterns that would be more tentative in each review taken on its own.

Technology Applications

The most important finding of the elementary and secondary reviews is the consistently strong impacts of applications of technology in science teaching. The studies tend to be small and to use measures made by the experimenters (to assess content taught equally in experimental and control groups), but nevertheless the impacts are impressive. They contrast strongly with findings from studies of technology applications in mathematics (Cheung & Slavin, 2013; Slavin et al., 2009a) and in reading (Cheung & Slavin, 2012; Slavin et al., 2009b).

A likely explanation for the different findings for science in contrast to math or reading is that technology is used very differently in science. In science, technology has been evaluated primarily as part of teachers' lessons to help students visualize science concepts. A good example is BrainPop, which provides cartoons to motivate and inform students about science ideas (Barak, Ashkar, & Dori, 2010). In contrast, technology applications in mathematics and reading tend to be drill-and-practice approaches designed to give students practice at their own level. These applications are generally disconnected from the teacher's instruction. The typical technology application in math or reading involves having students go to a computer lab or to the back of the class to work on individualized activities. None of the science technology programs that met the standards of this review operated in this way. As one point of interest, an elementary math program called *Time to Know* also gives teachers computerized content to use as part of whole-class instruction to help students visualize math concepts, and a small evaluation with fifth graders found this approach to be very effective (Rosen & Beck-Hill, 2012).

Another technology application with some evidence of effectiveness in science is providing all students with access to digital devices that link them to video, photos, illustrations, and additional explanations, as in the iBooks application created and evaluated by Pearson (Baughman et al., 2013). Finally, there was some evidence from brief studies reported in the elementary review that using technology to simulate laboratory exercises can be effective (Sun, Lin, & Wang, 2009; Sun, Lin, & Yu, 2008).

Nothing in this research on technology applications in science suggests that digital devices will, in themselves, enhance science learning, and there is nothing to suggest that approaches resembling computerized drill and practice are likely to make much of a difference. But the evidence that met the standards of this review provides reason for optimism about technology applications that help teachers increase the effectiveness of their lessons, especially in making concepts visual, motivating, and accessible. This evidence must be considered tentative, however, as the technology studies tended to use matched designs with small samples and often experimenter-made measures, all of which are associated with higher effect sizes.

Instructional Process Approaches

In both the elementary and the secondary science reviews, approaches emphasizing extensive professional development to help teachers implement well-defined classroom innovations had moderately positive effects on student learning. Innovations such as cooperative learning, metacognitive strategies, project-based inquiry, science-literacy integration, and teaching of science vocabulary, are very diverse, but generally improve student learning across a broad range of topics and age levels. Instructional process programs have consistently been the most effective programs in mathematics and reading, according to previous reviews using methods similar to those used in the present synthesis.

Science Kits

The greatest surprise of the elementary review was the consistent findings of near-zero impacts of programs providing students with sophisticated and comprehensive kits to help them carry out hands-on experiments. This included a large-scale study by Pine et al. (2006) comparing fifth grade classes using Insights, FOSS, and STC, widely known kits, in comparison to those using traditional textbooks. In addition to a test composed of items drawn from TIMSS, students took performance tests on topics they had not specifically studied but that should have registered general gains in scientific reasoning, involving carrying out four experiments: determining weight using a spring, testing the absorbency of different paper towels, comparing melting rates of ice cubes in salt vs. fresh water, and observing flatworms over three days. Students were individually observed doing these tasks by research assistants. Only the flatworm task showed significant differences, and the overall effect sizes were -0.02 for the TIMSS items and +0.11 for the performance measures, for a mean of +0.05.

Only one qualifying study, by Newman et al. (2012), evaluated a kit program in the elementary and middle grades, and none did so in high school. This study showed little impact of the approach (ES=+0.05).

The kit programs, usually developed under NSF funding, seem to embody the principles of inquiry teaching long advocated by science educators. They engage students in solving real science problems in the laboratory. Every science curriculum includes laboratory exercises, of course, but the kits enable teachers to make experiments the core of their teaching, in hopes that students will learn to think and act as scientists do and transfer principles they have enacted in the lab to broader understandings of how science works.

The disappointing findings of evaluations of kit programs in upper elementary and middle schools may suggest that science instruction at these levels needs more of a balance between teaching and laboratory work. Time for science teaching is limited, and a substantial focus on experiments means that other parts of the curriculum are not being adequately attended to. Students may learn the scientific method from a few experiences with high-quality, open-ended experiments. It is interesting to note that in this review and in the Slavin et al. (2014) elementary review, the programs that showed the strongest and most consistent impacts were ones that helped teachers do a better job of teaching all year, providing professional development and/or technology tools to increase teachers' ability to communicate essential ideas of science.

Textbooks

Science textbooks can be drivers of what gets taught in secondary science, but from the evidence summarized here, there is little reason to believe that it matters very much which textbooks teachers use. It is important to note that in every study, new textbooks were compared to existing textbooks, so the differences in what happened in classroom teaching may not have been great. The overall impact on student learning (weighted mean effect size =+0.10) is greater than zero, but there are clearly more effective approaches. It is interesting to note that in best-evidence syntheses of elementary and secondary reading and math programs, textbooks and other innovative curricula have also had near-zero impacts on achievement. There are many effective programs that do involve introducing new curriculum materials, but these programs also provide extensive professional development and create new models of classroom organization and instruction, and are therefore categorized as instructional process programs, not as textbook approaches.

Limitations

There are, of course, limitations to our findings which should be understood when considering the impact of this research. First, the current review only includes studies with rigorous research designs, such as quasi-experiments and randomized experiments with durations of at least 12 weeks. However, other research designs and studies with shorter durations are also valuable for theory building and concept testing.

The review also excludes studies that used experimenter-made measures of content taught in the treatment group, but not the control group. Such outcomes may be of importance to researchers or practitioners. Finally, although we carried out an extensive search for all potential studies using various databases and contacting developers and researchers, some studies may have been missed.

Conclusions

Science educators agree on the importance of inquiry in science education at all levels, and all science curricula include experiments to a greater or lesser degree. Neither inquiry nor experiments are matters of serious debate, though different educators and researchers do have different definitions of inquiry and different ideas of how it should be enacted in practice (Furtak et al., 2012; Minner et al., 2010; Schroeder et al., 2007). However, within this general consensus there remain essential questions about how to improve science achievement.

The findings of this review of research on middle and high school science programs are very consistent with those of an earlier review of elementary programs (Slavin et al., 2014). The types of programs that make a difference in student outcomes are those that help teachers teach more effective lessons: technology designed primarily to help students visualize science concepts, and instructional process models that provide teachers with extensive professional development to help them apply strategies such as cooperative learning, use of metacognitive skills, and science-literacy integration.

In contrast, approaches that attempt to improve science learning primarily through improving textbooks or providing teachers with kits to facilitate experiments have been less successful in rigorous evaluations.

What these findings imply is that teaching, not materials, is the core of the science classroom, and that investing in specific technologies and professional development designed to enhance the effectiveness of teaching is the best way forward in science education. More research and development are clearly needed, especially to build on the promising but still early-stage research on uses of technology to enhance teachers' lessons. As digital devices become universally available in science classes, these possibilities become ever more appealing and practicable. There are exciting possibilities in the research that suggest ways to accelerate students' science learning and reduce achievement gaps, but there is more we need to know about how to achieve these essential changes on a national scale.

References

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

Baughman, M., Ehmann, M., & Vilcheck, A. (2013). *The effects of iBooks Biology on student science achievement and motivation.* Retrieved March 10, 2015 from http://assets.pearsonschool.com/asset_mgr/current/201423/iBookBiologyWhitePapers.pdf

Cheung, A., & Slavin, R.E. (2012). How features of educational technology programs affect student reading outcomes: A meta-analysis. *Educational Research Review, 7*, 198-215.

Cheung, A., & Slavin, R. E. (2013a). Effects of educational technology applications on reading outcomes for struggling readers: A best-evidence synthesis. *Reading Research Quarterly, 48*, 277-299.

Cheung, A., & Slavin, R. E. (2013b). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review, 9,* 88-113.

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

Duschl, R.A. (2003). Assessment of inquiry. In J.M. Atkin & J. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41-59). Arlington, VA: NSTA Press.

Duschl, R.A. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of Research in Education, 32*, 268-291.

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.

Dunleavy, M., & Heinecke, W. F. (2008). The impact of 1:1 laptop use on middle school math and science standardized test scores. *Computers In The Schools, 24*, 7-22.

Eddy, R.M. & Berry, T. (2005). *The effects of Prentice Hall Biology on student performance: Pilot study*. Claremont, CA: Claremont Graduate University.

Eddy, R.M. & Berry, T. (2007). *A randomized control trial to test the effects of Prentice Hall's Miller and Levine (2006) Biology curriculum on student performance:* Claremont, CA: Claremont Graduate University.

Eddy, R. M., Ruitman, H. T., Sloper, M., & Hankel, N. (2010). *The effects of Miller & Levine Biology (2010) on student performance.* La Verne, CA: Cobblestone Applied Research and Evaluation.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching. *Review of Educational Research, 82,* 300-329.

Green, J. M. (2010). *Effects of peer-mediated vocabulary intervention on science achievement of seventh grade students with and without learning disabilities*. Unpublished doctoral dissertation, Clemson University.

Harris, C. J., Penuel, W. R., DeBarger, A., D’Angelo, C., & Gallagher, L. P. (2014). *Curriculum materials make a difference for next generation science learning: Results from year 1 of a randomized control trial.* Menlo Park, CA: SRI International.

Harwood, W. S., & McMahon, M. M. (1997). Effects of integrated video media on student achievement and attitudes in high school chemistry. *Journal of Research In Science Teaching, 34*, 617-31.

Heller, J. I. (2012). *Effects of Making Sense of SCIENCE[TM] professional development on the achievement of middle school students, including English language learners. Final report. NCEE 2012-4002*. Washington, DC: National Center For Education Evaluation and Regional Assistance.

Huffman, D., Goldberg, F., & Michlin, M. (2003). Using computers to create constructivist learning environments: Impact on pedagogy and achievement. *Journal of Computers in Mathematics and Science Teaching, 22,* 151–168. (452, 4520)

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

Lavoie, D.R. (1999). Effects of emphasizing hypothetico-predictive reasoning within the science learning cycle on high school students' process skills and conceptual understandings in biology. *Journal of Research in Science Teaching, 36*, 1127–1147.

Liao, Y. K. (1999). Effects of hypermedia on students' achievement: a meta-analysis. *Journal of Educational Multimedia and Hypermedia*, 8, 255-277.

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

Marx, R. W. (2012). Large-scale interventions in science education: The road to utopia? *Journal of Research on ScienceTeaching, 49,* 420–427.

Mevarech, Z. R., & Kramarski, B. (1997). IMPROVE: A multidimensional method for teaching mathematics in heterogeneous classrooms. *American Educational Research Journal, 34*, 365–394.

Michalsky, T. (2013). Integrating skills and wills instruction in self-regulated science text reading for secondary students. *International Journal of science education, 35,* 1846-1873.

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

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*, 474-496.

National Center for Education Statistics (2012). *The Nation's Report Card: Science 2011* (NCES 2012–465). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Research Council (1996). *National science education standards*. Washington, DC: National Academies Press.

National Research Council (2000). *Inquiry and the national science education standards: A guide for teaching and learning.* Washington, DC: National Academies Press.

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

Newman, D., Finney, P. B., Bell, S., Turner, H., Jaciw, A. P., Zacamy, J. L., & ... National Center for Education Evaluation and Regional Assistance. (2012). *Evaluation of the effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI). Final Report. NCEE 2012-4008.* Washington, DC: National Center For Education Evaluation And Regional Assistance.

Penuel, W. R., & Fishman, B. J. (2012). Large-scale intervention research we can use. *Journal of Research in Science Teaching, 49*, 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*, 467-484.

Prentice Hall (2003). *Prentice Hall Science 2000–2001 study comparing performance of students using Prentice Hall Science Explorer with students using other science programs.* Upper Saddle River, NJ: Author.

Resendez, M. & Azin, M. (2006). *2005 Prentice Hall Science Explorer randomized control trial.* Jackson, WY: PRES Associates, Inc.

Resendez, M. & Azin, M. (2013). *A study on the effects of Houghton Mifflin Harcourt's Science Fusion: Year 2 comprehensive final report*. Jackson, WY: PRES Associates, Inc.

Resendez, M., DuBose, D. & Azin, M. (2011). *A study on the effects of Pearson's Interactive Science 2011 Program.* Jackson, WY: PRES Associates, Inc.

Rosen, Y., & Beck-Hill, D. (2012). Intertwining digital content and a one-to-one laptop environment in teaching and learning: Lessons from the *Time To Know* program. *Journal of Research on Technology in Education, 44*, 225-227.

Schroeder , C., Scott, T., Tolson , H., Huang , T., & Lee, Y. (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*, 1436-1460.

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

Shannon, L. & Grant, B. (2012). *A final evaluation report of Houghton Mifflin Harcourt's Holt McDougal Biology*. Charlottesville, VA: Magnolia Consulting, LLC.

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

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

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

Slavin, R. E., Lake, C., Hanley, P., & Thurston, A. (2014). Experimental evaluations of elementary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching, 51*, 870-901.

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

Slavin, R.E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best evidence synthesis. *Reading Research Quarterly, 43*, 290-322.

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

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*, 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.

Taylor, J, Getty, S, Kowalski, C, Wilson, J, Carlson, J, & Van Scotter, P (2014). *An efficacy trial of research-based curriculum materials with curriculum-based professional development*. Article submitted for publication.

Turpin, T. J. (2000). *A study of the effects of an integrated, activity-based science curriculum on student achievement, science process skills, and science attitudes*. Unpublished doctoral dissertation, University of Louisiana at Monroe.

Table 1
Instructional Process Approaches

Study	Design	Duration	N	Grades, Subjects	Sample Characteristics	Posttest	Effect Size	Overall Effect Size
Peer-Mediated Vocabulary Intervention								
Green (2010)	Randomized Quasi-Experiment	4 months	675 students (311E, 364C)	7 Science	2 adjacent school districts in a Southeastern state, US	Science Assessment	+0.24	+0.24
IMPROVE								
Michalsky (2013)	Randomized Quasi-Experiment	12 weeks	4 teachers 95 students (49E, 46C)	10 Biology	Israel	Science Literacy Items from PISA	+1.26	+1.26
BSCS Inquiry Approach								
Taylor, Getty, Kowalski, Wilson, Carlson, & Scottier (2014)	Cluster Randomized	2 years	18 schools (9E, 9C) 3052 students (1509E, 1543C)	10-11 Integrated Science	Suburban/rural WA. 45% FL 52%W, 27%H, 8%A, 7% AA	Washington State Science Assessment	+0.09	+0.09
Project-Based Inquiry Science								
Harris, Penuel, DeBarger, D'Angelo, & Gallagher (2014)	Cluster Randomized	1 year	94 teachers (55E, 39C) 2400 students	6 Science	42 schools in one large urban highly diverse school district	Physical Science (Energy)	+0.21	+0.23
						Earth Science	+0.25	

Prediction/Discussion-Based Learning Cycle Instruction								
Lavioe (1999)	Matched	3 months	251 students (131E, 120C)	10 Biology	Low to middle SES schools	Group Assessment of Logical Thinking	+0.42	+0.46
						Processes Biological Investigation Test	+0.40	
						Concept Understanding Test	+0.56	
Making Sense of SCIENCE								
Heller (2012)	Cluster Randomized	1 year	181 teachers (90E, 91C) 7000 students	8 Physics	Diverse schools from 55 districts in California and Arizona	ATLAS Test of Force and Motion	+0.11	+0.07
						California Standards Test--Physical Science	+0.03	
Integrated, Activity-Based Science Curriculum								
Turpin (2000)	Matched	1 year	14 classes 982 students (532E, 450C)	7 Integrated Science	Low SES schools in a mid-size central city in Louisana	ITBS	+0.16	+0.16

Table 2
Technology Approaches

Study	Design	Duration	N	Grades, Subjects	Sample Characteristics	Posttest	Effect Size	Overall Effect Size
BrainPOP								
SEG (2009)	Matched	4 months	128 students (59E, 69C)	8 Science	Schools in Palm Beach, FL and New York City	SAT Science	+0.41	+0.41
Constructing Physics Understanding (CPU)								
Huffman, Goldberg, & Michlin (2003)	Matched	1 year	13 teachers (8E, 5C) 194 students (116E, 78C)	High School Physics	No information	Force Concept Inventory Test	+0.64	+0.64
Integrated Video Media								
Harwood & McMahan (1997)	Matched	1 year	18 classes 373 students (182E, 191C)	9-12 Chemistr y	Students from a multiculturally diverse metropolitan region of the East Coast	High School Subject Test	+0.71	+0.71
iBooks								
Baughman, Ehmann, & Vilcheck (2013)	Matched	1 year	3 teachers 6 classes 178 students (83E, 95C)	High School Biology	Suburban NJ, Charter in SC 12% FL, 70%W, 9%H, 8%AA, 4% Asian	SAT 10 Science	+0.25	+0.25
1:1 Laptop Use								
Dunleavy & Heinecke (2007)	Cluster Randomized	2 years	14 classes 163 students (52E, 111C)	Middle School Science	Middle school students in an urban school district in a Mid-Atlantic state	Science Standardized Achievement Test	+0.24	+0.24

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

Table 3
Science Kits

Study	Design	Duration	N	Grades, Subjects	Sample Characteristics	Posttest	Effect Size	Overall Effect Size
Alabama Math, Science, and Technology Initiative (AMSTI)								
Newman et al. (2012)	Cluster Randomized	1 year	79 schools (39E, 40C) 780 teachers (102E, 90C) 7528 students (4082E, 3446C)	5, 7 Science	Schools from 5 regions of Alabama	SAT 10 Science Assessment	+0.05	+0.05

Table 4
Textbooks

Study	Design	Duration	N	Grades, Subjects	Sample Characteristics	Posttest	Effect Size	Overall Effect Size
Miller & Levine Biology								
Eddy & Berry (2005)	Matched	1 year	4 classes 205 students (92E, 113C)	9-10 Biology	Ethically diverse high school in California	Standards-based biology assessment	+0.01	+0.01
Eddy & Berry (2007)	Cluster Randomized	1 year	16 teachers 1108 students (541E, 567C)	9-12 Biology	5 high schools across 4 states	Biology Test	+0.02	+0.02
Eddy, Ruitman, Sloper, & Hankel (2010)	Cluster Randomized	1 year	24 teachers 1974 students (1126E, 848C)	9-10 Biology	6 high schools in 5 states from suburban and rural areas	SAT 9	+0.18	+0.10
						Biology Core Assessment	+0.02	
Holt McDougal Biology								
Shannon & Grant (2012)	Cluster Randomized	1 year	24 teachers 1255 students (671E, 584C)	9-10 Biology	8 schools in 7 districts	SAT 10 Science Assessment	+0.06	+0.09
						PASS Biology Achievement	+0.12	
Prentice Hall Science Explorer								
Prentice Hall (2003)	Matched	1 year	12 classes 223 students (108E, 115C)	8 Science	Six schools in four states (CO, NJ, WA, & WI)	TerraNova CTBS Basic Battery Plus	+0.12	+0.12

Resendez & Azin (2006)	Cluster Randomized	1 year	17 teachers 1255 students (646E, 619C)	6-8 Science	4 geographically dispersed schools	ITBS	-0.04	+0.04
						TIMSS	+0.12	
Pearson Interactive Science Program								
Resendez, DuBose, & Azin (2011)	Cluster Randomized	1 year	35 teachers 1362 students (634E, 728C)	6-8 Science	9 geographically dispersed schools	TerraNova Science	+0.14	+0.14
Houghton Mifflin Science Fusion								
Resendez & Azin (2013)	Cluster Randomized	2 years	27 classes 576 students (263E, 313C)	6-7 Science	3 schools	ITBS	+0.39	+0.39