Abstract Title Page

Title: Effective Programs for Elementary Science: A Best-Evidence Synthesis

Authors/Affiliations:

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

Cynthia Lake
Johns Hopkins University

Pam Hanley
University of York

Allen Thurston
Durham University
Effective Programs for Elementary Science:  
A Best-Evidence Synthesis

Background/Context
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), there has been general agreement in the U.S. about what students should learn in science, and a consensus that science should be taught using inquiry-oriented methods that emphasize conceptual understanding rather than just facts. Yet beyond this broad agreement, what do we know about what works in elementary science? While there have been several reviews of research on various aspects of science teaching, there has not been a comprehensive review of evaluations of alternative approaches to science education at any instructional level. The U.S. Department of Education’s What Works Clearinghouse, which has been reviewing research on educational programs for eight years, does not even have science in its long-term plans.

There have been several reviews of research on various aspects of science education, such as inquiry teaching (Anderson, 2002; Bennett, Lubben, & Hogarth, 2006; Minner, Levy, & Century, 2010), small-group methods (Bennett, Lubben, Hogarth, & Campbell, 2004; Lazarowitz & Hertz-Lazarowitz, 1998), and overall methods (Hopkins et al., 2002; Fortus, 2008; Gray, 2005; Scott et al., 2005). Most of these reviews include elementary as well as secondary studies but secondary studies are far more common. The only review of the research that focuses solely on elementary science is an unpublished paper written for Alberta (Canada) school leaders (Gustafson, MacDonald, & d’Entremont, 2007).

Purpose
This paper reports the findings of a systematic review of research on elementary science programs.

Research Design
The review methods for elementary science applied in this paper are similar to those used in math by Slavin & Lake (2008) and Slavin, Lake, & Groff (2009), and in reading by Slavin, Lake, Chambers, Cheung, & Davis (2009). These reviews used an adaptation of a technique called best evidence synthesis (Slavin, 2008), which seeks to apply consistent, well-justified standards to identify unbiased, meaningful information from experimental studies, discuss each study in some detail, and pool effect sizes across studies in substantively justified categories. Best-evidence syntheses are similar to meta-analyses (Cooper, 1998; Lipsey & Wilson, 2001), adding an emphasis on narrative description of each study’s contribution and limiting the review to studies meeting the established criteria. They are also similar to the methods used by the What Works Clearinghouse (2009).
Literature Search Procedures

A broad literature search was carried out in an attempt to locate every study that could possibly meet the inclusion requirements. Electronic searches were made of educational databases (JSTOR, ERIC, EBSCO, Psych INFO, Dissertation Abstracts) using different combinations of key words (for example, “elementary students” and “science achievement”) and the years 1970-2011. Results were then narrowed by subject area (for example, “educational software,” “science achievement,” “instructional strategies”). In addition to looking for studies by key terms and subject area, we conducted searches by program name. Web-based repositories and education publishers’ websites were examined. We attempted to contact 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. Studies that meet an initial screen for germaneness (i.e., they involved elementary science) and basic methodological characteristics (i.e., they had a well-matched control group and a duration of at least 4 weeks) were independently read and coded by at least two researchers. Any disagreements in coding were resolved by discussion, and additional researchers were asked to read any articles on which there remained disagreements.

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

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.
2. The studies involved approaches that began when children were in grades K-5, plus sixth graders if they were in elementary schools.
3. The studies compared children taught in classes using a given science program or practice with those in control classes using an alternative program or standard methods. The program or practice had to be one that could, in principle, be used in ordinary science classes.
4. Studies could have taken place in any country, but the report had to be available in English.
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 are no indications of initial inequality. Studies with pretest differences of more than 50% of a standard deviation were excluded because, even with analyses of covariance, large pretest differences cannot be adequately controlled for, as underlying distributions may be fundamentally different (Shadish, Cook, & Campbell, 2002). Studies using pretests with indications of ceiling or floor effects were excluded.
7. The dependent measures included quantitative measures of science performance. Experimenter-made measures were accepted if they covered content taught in control as well
as experimental groups, but measures of science objectives inherent to the program (and unlikely to be emphasized in control groups) were excluded.

8. A minimum study duration of 4 weeks was required.

9. Studies had to have at least two teachers and 15 students in each treatment group.

**Findings**

The most important finding of the present review is the very limited number of rigorous evaluations of elementary science programs. After an exhaustive search involving examination of hundreds of published and unpublished articles that have appeared since 1970, only 13 elementary science studies met the review standards. All of these involved children in grades 3-6. 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), and a similar review of middle and high school mathematics studies identified 100 qualifying studies (Slavin et al., 2009).

The few elementary science studies that did meet the inclusion criteria provide useful information on several approaches to improving outcomes in science teaching. The studies can be grouped into three categories of treatments. One focuses on *instructional processes* for teachers, especially on inquiry teaching, cooperative learning, and integrating science and reading. The theory of action unifying this category of approaches is an emphasis on teachers improving science learning by using specific, well-articulated strategies in the classroom. These interventions invariably emphasize professional development and coaching to help teachers use promising approaches.

A second category of interventions also focuses on instructional processes, but in addition they provide teachers with kits and specific guidelines for hands-on inquiry-oriented explorations. The theory of action underlying these approaches 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, outcomes will improve. An example of this approach is science kits such as FOSS and STS. These provide 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.

A third type of approach is emphasizes the use of *technology* to enhance student outcomes. This may be individual technologies, such as computer-assisted instruction, as well as class-focused technology, such as video and interactive whiteboard technologies.

A surprising finding from the largest and best-designed of the studies is the limited achievement impact of elementary science programs that provide teachers with kits to help them make regular use of hands-on, inquiry-oriented activities. These include evaluations of the well-regarded FOSS, STC, and Teaching SMART programs, none of which showed positive achievement impacts. One might argue that traditional science tests might not be sensitive to the more sophisticated understandings of scientific process that are the targets of these inquiry-oriented approaches, but studies by Pine et al. (2006) and K. Borman et al. (2009) used well-designed process measures as well as multiple choice assessments, and failed to find any 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).

In contrast, several equally inquiry-oriented professional development programs that do 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; Scott, 2005), cooperative learning (Baines et al., 2007; Jang, 2010), science-reading integration (Romance & Vitale, 1992, 2011), science vocabulary instruction (Rosebrock, 2007), all of these studies found significant positive effects of inquiry-oriented professional development on conventional measures of science achievement. Further, a study of the use of humorous, explanatory videos (but not science kits) also found positive outcomes (Barak et al., 2010).

Conclusions

Although the limited number of qualifying studies makes explanations of these divergent outcomes tentative at best, it is nevertheless interesting to speculate about their meaning. How could the provision of science kits carefully designed to facilitate hands-on inquiry reduce the effectiveness of science interventions?

One possible answer may lie in the nature of practical science teaching in elementary schools. In reality, time and resource limitations make it difficult to cover the entire science curriculum. Elementary teachers who spend a great deal of time on laboratory exercises may be taking time away from coverage of the rest of the curriculum, objectives not covered using the kits by an individual teacher. Further, professional development targeted toward helping teachers use kits may not help them enhance their effectiveness on the units taught without kits.

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

If this explanation turns out to be correct, it suggests that elementary science programs might enhance their effectiveness on broadly focused measures of science learning by providing teachers with professional development on all objectives, both those emphasized in the kits and those directly taught to students. That is, the findings of the qualifying studies do not necessarily 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; Bennett, Lubben, & Hogarth, 2006; Anderson, 2002). Yet few if any elementary science teachers use hands-on inquiry activities every day to cover all of the curricular expectations in today’s state and national standards. In order to make a substantial difference on broad measures of science learning, teachers may need more effective pedagogical strategies for all objectives and all teaching modes.

Far more research and development are needed to identify effective and replicable approaches to improving science achievement outcomes for elementary schools. Science education needs to move beyond brief and artificial pilot tests of exciting new methods and technologies to put them to the test in real schools over extended time periods with valid and comprehensive measures of what students should know and be able to do in science. Science education researchers need to use the tools of science to evaluate and progressively improve the programs and practices needed to help elementary teachers build a scientifically literate society.
Appendix A. References
References are to be in APA version 6 format.


### Appendix B. Tables and Figures

#### Table 1: Instructional Process Programs Without Science Kits

<table>
<thead>
<tr>
<th>Study</th>
<th>Design</th>
<th>Duration</th>
<th>N</th>
<th>Grade</th>
<th>Sample Characteristics</th>
<th>Posttest</th>
<th>Effect Sizes by Subgroup/Measure</th>
<th>Overall Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Increasing Conceptual Challenge</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mant, Wilson, &amp; Coates (2007)</td>
<td>Matched</td>
<td>1 year</td>
<td>32 schools (16E, 16C) 1120 students (560E, 560C)</td>
<td>Year 6 10-11 yrs old</td>
<td>Rural and village schools in Oxfordshire, England, mostly White, not low SES</td>
<td>(National) Key Stage 2 science tests</td>
<td></td>
<td>+0.33</td>
</tr>
<tr>
<td><strong>SPRinG - Cooperative Learning</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baines, Blatchford, &amp; Chowne (2007)</td>
<td>Matched</td>
<td>2 years</td>
<td>31 schools (12E, 19C) 51 classes (21E, 40C) 1587 students (560E, 1027C)</td>
<td>Years 4-5 8-10 yrs old</td>
<td>Schools in London, England</td>
<td>Items adapted from standardized tests for Year 6, simplified for younger children</td>
<td></td>
<td>+0.21</td>
</tr>
<tr>
<td><strong>Science IDEAS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romance &amp; Vitale (2011)</td>
<td>Matched</td>
<td>6 years</td>
<td>24 schools (12E, 12C)</td>
<td>3-8</td>
<td>Large urban district in Florida 40% FL 47%W, 29%AA, 19%H</td>
<td>ITBS Science</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romance &amp; Vitale (1992)</td>
<td>Matched</td>
<td>1 year</td>
<td>7 classes (3E, 4C) 128 students (51E, 77C)</td>
<td>4</td>
<td>Large urban district in Florida</td>
<td>MAT Science</td>
<td>ITBS Reading</td>
<td>+0.90</td>
</tr>
<tr>
<td><strong>Collaborative Concept-Mapping with Co-Teaching</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jang (2010)</td>
<td>Matched</td>
<td>8 weeks</td>
<td>114 students (58E, 56C)</td>
<td>4</td>
<td>Science classes in Taiwan</td>
<td>Schoolwide uniform science test</td>
<td></td>
<td>+0.54</td>
</tr>
<tr>
<td>Study</td>
<td>Method</td>
<td>Duration</td>
<td>Sample Size</td>
<td>Matched Groups</td>
<td>Controls</td>
<td>Intervention Duration</td>
<td>Controls</td>
<td>Intervention Effect</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>--------</td>
<td>----------</td>
<td>-------------</td>
<td>----------------</td>
<td>----------</td>
<td>------------------------</td>
<td>----------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Rosebrock (2007)</td>
<td>Matched</td>
<td>12 weeks</td>
<td>686 students (401E, 205C)</td>
<td>Matched on test scores but not demographics</td>
<td>5</td>
<td>Middle-class suburb of Houston</td>
<td>TAKS Earth and Space Science Subtest</td>
<td>+0.24</td>
</tr>
<tr>
<td>Scott (2005)</td>
<td>Matched</td>
<td>1 year</td>
<td>99 students (66E, 33C)</td>
<td>3</td>
<td>Large, diverse district outside of Houston, TX 54%H, 37%AA, 5%W, 83% FL, 40% LEP</td>
<td>ITBS-Science</td>
<td>+0.29</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Duration</td>
<td>N</td>
<td>Grade</td>
<td>Sample Characteristics</td>
<td>Posttest</td>
<td>Effect Sizes by Subgroup/Measure</td>
<td>Overall Effect Size</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>---------</td>
<td>----------</td>
<td>---------</td>
<td>-------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>---------------------------</td>
<td>-------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td><strong>Insights, FOSS, and STC</strong></td>
<td>Matched</td>
<td>1 yr</td>
<td>41 classrooms</td>
<td>5</td>
<td>9 school districts in CA, AZ, and NV Compared high SES group to low SES group for hands-on curricula vs traditional textbooks 500 students in each group</td>
<td>TIMSS</td>
<td>-0.02</td>
<td>+0.05</td>
</tr>
<tr>
<td>Pine, Aschbacher, Roth, Jones, McPhee, Martin, Phelps, Kyle, &amp; Foley (2005)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance tasks</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FOSS</strong></td>
<td>Random</td>
<td>14 weeks</td>
<td>5 classes (2E, 3C) 103 students (38E, 65C)</td>
<td>5</td>
<td>Urban district in TX, 49% minority</td>
<td>CTBS science</td>
<td>+0.29</td>
<td>+0.48</td>
</tr>
<tr>
<td>Leach (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Electricity and magnetism test</td>
<td>+0.67</td>
<td></td>
</tr>
<tr>
<td><strong>System-Wide Change for All Learners and Educators (SCALE)</strong></td>
<td>Cluster Random</td>
<td>1 yr</td>
<td>71 schools (33E, 38C)</td>
<td>4, 5</td>
<td>Los Angeles USD 73%H, 11%W, 8%AA 76% FL, 33%ESL</td>
<td>LAUSD science assessments</td>
<td>-0.11</td>
<td></td>
</tr>
<tr>
<td>G. Borman, Gamoran, &amp; Bowdon (2008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Life science</td>
<td>-0.27</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Earth science</td>
<td>+0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Physical science</td>
<td>-0.08</td>
<td></td>
</tr>
<tr>
<td><strong>Teaching SMART</strong></td>
<td>Cluster Random</td>
<td>3 yrs</td>
<td>20 schools</td>
<td>3-5</td>
<td>Pasco County, FL</td>
<td>PASS</td>
<td></td>
<td>+0.04</td>
</tr>
<tr>
<td>K. Borman, Boydston, Lee, Lanehart, &amp; Cotner (2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Multiple choice</td>
<td>+0.08</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Performance</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>
## Science Curriculum Improvement Study (SCIS)

<table>
<thead>
<tr>
<th>Study</th>
<th>Random</th>
<th>Duration</th>
<th>Participants</th>
<th>Age</th>
<th>Location</th>
<th>Test</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hendricks (1978)</td>
<td>Random</td>
<td>12 weeks</td>
<td>308 students (168E, 140 C)</td>
<td>5</td>
<td>2 rural counties in Southwest Mississippi (91% poverty)</td>
<td>STEP Science Test</td>
<td>+0.14</td>
</tr>
<tr>
<td>Study</td>
<td>Design</td>
<td>Duration</td>
<td>N</td>
<td>Grade</td>
<td>Sample Characteristics</td>
<td>Posttest</td>
<td>Effect Sizes by Subgroup/Measure</td>
</tr>
<tr>
<td>--------------</td>
<td>--------</td>
<td>----------</td>
<td>--------------------</td>
<td>-------</td>
<td>------------------------</td>
<td>---------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>Brain Pop</td>
<td>Matched</td>
<td>1 yr</td>
<td>7 schools (5E, 2C)</td>
<td>4, 5</td>
<td>Students in Israel</td>
<td>Measure based on Israeli national standards</td>
<td>+0.43</td>
</tr>
</tbody>
</table>