Experimental and Quasi-Experimental Studies of Inquiry-Based Science Teaching: A Meta-Analysis
Erin Marie Furtak, Tina Seidel, Heidi Iverson and Derek C. Briggs

REVIEW OF EDUCATIONAL RESEARCH 2012 82: 300 originally published online 8 August 2012
DOI: 10.3102/0034654312457206

The online version of this article can be found at:
http://rer.sagepub.com/content/82/3/300
Experimental and Quasi-Experimental Studies of Inquiry-Based Science Teaching: A Meta-Analysis

Erin Marie Furtak  
University of Colorado at Boulder

Tina Seidel  
TUM School of Education

Heidi Iverson  
University of Colorado Denver

Derek C. Briggs  
University of Colorado at Boulder

Although previous meta-analyses have indicated a connection between inquiry-based teaching and improved student learning, the type of instruction characterized as inquiry based has varied greatly, and few have focused on the extent to which activities are led by the teacher or student. This meta-analysis introduces a framework for inquiry-based teaching that distinguishes between cognitive features of the activity and degree of guidance given to students. This framework is used to code 37 experimental and quasi-experimental studies published between 1996 and 2006, a decade during which inquiry was the main focus of science education reform. The overall mean effect size is .50. Studies that contrasted epistemic activities or the combination of procedural, epistemic, and social activities had the highest mean effect sizes. Furthermore, studies involving teacher-led activities had mean effect sizes about .40 larger than those with student-led conditions. The importance of establishing the validity of the treatment construct in meta-analyses is also discussed.

Keywords: science teaching, science learning, meta-analysis, inquiry-based teaching, student achievement.

In the past 50 years, major educational policy organizations in the United States and across the world have emphasized that students learn by engaging in the thinking processes and activities of scientists (e.g., American Association for the
Meta-Analysis of Inquiry-Based Teaching

Advancement of Science, 1993; Mullis, Martin, Ruddock, O’Sullivan, & Preuschoff, 2009; National Research Council [NRC], 2001; Organisation for Economic Co-Operation and Development, 2009). This approach has often been described as inquiry-based teaching, or science as inquiry, and includes students drawing upon their scientific knowledge to ask scientifically oriented questions, collect and analyze evidence from scientific investigations, develop explanations of scientific phenomena, and communicate those explanations with their teacher and peers (NRC, 1996). Although inquiry has been emphasized as a popular reform in science education since the 1960s, it returned to prominence in the 1990s with the National Science Education Standards (NRC, 1996, 2001).

Despite this emphasis, the efficacy of inquiry-based teaching has been continually challenged (Kirschner, Sweller, & Clark, 2006; Mayer, 2004). Critics of inquiry-based teaching have argued that its minimally guided approach does not provide sufficient structure to help students learn the important concepts and procedures of science. These critics often characterize the inquiry-oriented teacher as staying in the background while students engage in self-guided, hands-on activities of dubious value (e.g., Kirschner et al., 2006). These critics have advocated for traditional, direct instruction in which teachers deliver content to students through carefully designed lectures and verification-style laboratory activities.

While the debate between inquiry-based and traditional instructional approaches has continued to simmer, researchers have investigated inquiry-based teaching reforms with particular interest in the specific features that appear to lead to increased student learning. These studies typically have an experimental or quasi-experimental design in which student outcomes are compared across two groups of students—a control group taught in a traditional lecture-based manner and a treatment group taught with some form of inquiry-based teaching. These studies assume that if inquiry-based teaching is indeed a more effective way to help students learn, students in the treatment conditions will outperform students in the control conditions on measures of conceptual understanding. In fact, a number of meta-analyses of reforms in science education have come to this conclusion (Bredderman, 1983; Schroeder, Scott, Tolson, Huang, & Lee, 2007; Shymansky, Hedges, & Woodworth, 1990; Weinstein, Boulanger, & Walberg, 1982).

Despite these findings, critiques of inquiry-based teaching have persisted, in part because of disagreements among researchers about what features define this instructional approach (e.g., Furtak, Shavelson, Shemwell, & Figueroa, 2012; Hmelo-Silver, Duncan, & Chinn, 2007). Inquiry has been defined in a number of ways, ranging from simple descriptions of students actively guiding their own learning with the teacher acting as facilitator to more elaborated lists of actions for the teacher, student, and curriculum (e.g., NRC, 1996, 2001). This situation is further complicated by the fact that the field of science education in the United States has moved away from using the term inquiry as the primary descriptor for science education reform and instead has focused on articulating how students can engage in science as a body of knowledge, a set of practices, and a process of participation in those practices (NRC, 2007). Indeed, the recently released Framework for K–12 Science Education focuses on identifying these practices, as well as crosscutting themes and core ideas for science education (Board on Science Education [BOSE], 2012).
The ways in which inquiry has been defined has consequences for the validity of meta-analyses that attempt to synthesize the causal inferences made by individual studies. For example, one study might define inquiry as students collecting data in response to a question their teacher had posed and contrast this with a condition in which students analyze data provided by their teacher. A second study might have the teacher guiding students to develop explanations about a scientific phenomenon and contrast this condition with students who also developed explanations, but with the support of an online tool. Although both studies belong within the domain of inquiry-based teaching, the activities and cognitive processes in which students engage are quite different. In addition, the role of the teacher varies between the conditions. These variations in the way that inquiry-based teaching is defined have consequences for the inferences made in research syntheses about the effectiveness of the approach (Briggs, 2008).

This purpose of this article is to reevaluate the effect of inquiry-based reforms on student learning of science in the 10 years (1996–2006) immediately following the publication of the National Science Education Standards (NRC, 1996). To do so, we will apply a framework for inquiry-based reforms based on two dimensions: (a) the cognitive domains of inquiry in which students are engaged and (b) the extent to which activities are guided by the teacher or student. These distinctions will be made for both the experimental and control groups.

Previous Inquiry Meta-Analyses

In past years, several meta-analyses have been conducted on studies of inquiry-based science teaching. These studies fall roughly into two categories: those that focused explicitly on the outcomes of reform-oriented versus traditional curricula and those that contrasted inquiry-based teaching with other approaches. These studies generally computed effect sizes by taking the difference between mean achievement outcomes for treatment and control groups and dividing by the standard deviation of the control group (e.g., Glass’s Δ; Glass, 1976). The following section will briefly review these meta-analyses for the purpose of identifying how each one conceptualized inquiry-based teaching.

A set of meta-analyses conducted in the 1980s framed inquiry-based teaching as innovative (Weinstein, 1982), activity based (Bredderman, 1983), process oriented (Shymansky, Hedges, & Woodworth, 1990), and discovery oriented (Wise & Okey, 1983). Table 1 summarizes their mean effect sizes. Three studies contrasted reform-oriented with traditional curricula, presenting mean effect sizes for several types of outcomes (Bredderman, 1983; Shymansky et al., 1990; Weinstein et al., 1982). In addition, two meta-analyses coded subsets of studies that explored the extent to which the teacher or the student guided the activity and the level of inquiry (Lott, 1983; Wise & Okey, 1983). Mean effect sizes across these studies ranged from .06 to .35, with larger effects found for inquiry-discovery and guided exploration as compared to teacher direction and structured guidance (Lott, 1983; Wise & Okey, 1983). Medium-level inquiry had a larger effect on student outcomes than low-level inquiry (Lott, 1983).

The most recently published meta-analysis on science teaching, conducted by Schroeder et al. (2007), reviewed experimental and quasi-experimental studies.
Their coding system included inquiry strategies that they defined as less recipe-style, more student-centered instruction in which students answer scientifically oriented research questions. The authors found an effect size of .65 for this subset of studies, a number that exceeds those found in the 1980s meta-analyses.

Although not a meta-analysis, the Inquiry Synthesis Project is the most recent effort to survey the results of studies of inquiry-based teaching (Minner, Levy, & Century, 2010). This project included 138 studies and did not exclude studies on the basis of design. As a result, unlike the preceding syntheses, their final analysis merged results of qualitative studies with experimental, quasi-experimental, and nonexperimental studies. Studies were coded according to design, methodological rigor, inquiry saturation, and impact. Within their larger sample, Minner et al. (2010) identified a subset of 42 comparative studies that contrasted inquiry-based teaching with other approaches. Of these studies, 55% found students in the condition with “higher inquiry saturation” to outperform students in the comparison group (Minner at al., 2010, p. 483). However, the authors did not calculate an effect size for the subset of these studies. Additionally, the authors also found that six of the subset of nine studies that directly studied student versus teacher direction found a “statistically significant increase in student conceptual learning when there was more student responsibility in the instruction (and higher inquiry satisfaction)” (Minner at al., 2010, p. 492). Overall, the authors argued that their synthesis indicates a “clear, positive trend” in favor of inquiry-based teaching” (p. 474).

<table>
<thead>
<tr>
<th>Study</th>
<th>Contrast</th>
<th>N studies</th>
<th>Effect size estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weinstein, Boulanger, and Walberg (1982)</td>
<td>Innovative versus traditional curricula</td>
<td>151</td>
<td>.31</td>
</tr>
<tr>
<td>Bredderman (1983)</td>
<td>Activity-based versus non-activity-based elementary curricula</td>
<td>57</td>
<td>.35</td>
</tr>
<tr>
<td>Wise and Okey (1983)</td>
<td>Total</td>
<td>309</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>Inquiry-discovery</td>
<td>38</td>
<td>.41</td>
</tr>
<tr>
<td></td>
<td>Teacher direction</td>
<td>28</td>
<td>.18</td>
</tr>
<tr>
<td>Lott (1983)</td>
<td>Composite</td>
<td>39</td>
<td>.06</td>
</tr>
<tr>
<td></td>
<td>Structured</td>
<td>8</td>
<td>-.14</td>
</tr>
<tr>
<td></td>
<td>Guided exploration</td>
<td>15</td>
<td>.43</td>
</tr>
<tr>
<td></td>
<td>Low level inquiry</td>
<td>12</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>Medium level inquiry</td>
<td>5</td>
<td>.41</td>
</tr>
<tr>
<td>Schroeder, Scott, Tolson, Huang, and Lee (2007)</td>
<td>Inquiry strategies</td>
<td>12</td>
<td>.65</td>
</tr>
</tbody>
</table>

a. Weighted.
b. Unweighted.
c. Mean of both weighted and unweighted effect sizes.
Motivation for a New Meta-Analysis

The preceding summary illustrates the variability in the ways that inquiry-based teaching has been defined and coded, something that has a direct bearing on a meta-analysis’s construct validity. Shadish, Cook, and Campbell (2002) defined construct validity as “the validity of inferences about the higher order constructs that represent sampling particulars across persons, treatments, outcomes and settings” (p. 17). In other words, the generalizability of the inferences one can make after combining effect sizes in a meta-analysis will be sensitive to the way that the sample of students has been selected, the way that the outcome variable has been measured, and the way that the treatment under investigation has been defined. It is important for meta-analysts to pay close attention to these implicit choices because when they have been carefully coded they can be used to help explain the variability in effect sizes from study to study. We argue that insufficient attention has been given to the operationalization of the inquiry construct in the case of prior meta-analyses of inquiry-based teaching and that this has masked important differences in the efficacy of distinct features of this instructional approach.

The majority of previous meta-analyses have relied upon expansive definitions of inquiry-based teaching (e.g., Bredderman, 1983; Shymansky et al., 1990; Weinstein et al., 1982). Although some have explicitly defined some aspects of inquiry teaching (e.g., Lott, 1983; Wise & Okey, 1983), others are much less clear about the parts that make up the whole (Schroeder et al., 2007). We argue that coding inquiry as a dichotomy, as opposed to existing on a spectrum, fails to capture the range of activities and thinking processes in which students might be engaged (Hmelo-Silver et al., 2007). The problem with using inquiry as a “black box” category is that it does not allow for distinctions between activities that are guided more by the teacher and those guided more by the student (NRC, 2001). Only a few of the prior reviews explore the amount of guidance provided by the teacher as a dimension of the reform being enacted in the study (Lott, 1983; Minner et al., 2010; Wise & Okey, 1983).

The shortcomings in the conceptualization of inquiry-based teaching in previous meta-analyses call into question the generalizability of claims being made about the effectiveness of inquiry as an instructional approach. In order to move research forward with regard to reforming traditional teaching, we need to know more about the specific effects of inquiry-based teaching on student learning by better articulating a model for this instructional approach. The following section will present a two-dimensional framework that we will use in this meta-analysis to compare and contrast the effects of inquiry-based teaching on student learning that have been found in the research literature.

Dimensions of Inquiry-Based Teaching

Part of the disagreement surrounding the effectiveness of inquiry-based teaching may be that the term inquiry itself has taken on multiple meanings in the science education reform literature. Inquiry has been alternatively used to describe (a) scientific ways of knowing (i.e., the work that scientists do), (b) a way for students to learn science, (c) an instructional approach, and (d) curriculum materials (Furtak et al., 2012). Some in the science education community have even argued that there is an inquiry threshold below which educational interventions can no longer
be identified as inquiry based (Minner et al., 2010). Given this, it is important to be precise and transparent about how the construct of inquiry is to be defined when conducting a meta-analysis. In this study, we define inquiry in terms of two dimensions: the cognitive and social activities of the student and the guidance provided to students by their teacher, their peers, or curriculum.

**Cognitive Dimension of Inquiry**

For the cognitive dimension of inquiry, we draw upon Duschl’s (2003, 2008) conceptualization of inquiry-based teaching and learning, which was developed on the basis of research and scholarly findings from cognitive and social psychology, the history and philosophy of science, and educational research. Duschl identified three categories of inquiry that included conceptual structures and cognitive processes that are used during scientific reasoning, epistemic frameworks used when scientific knowledge is developed and evaluated, and social interactions that shape how knowledge is communicated, represented, argued and debated.

The conceptual domain consists of the facts, theories, and principles of science, that is, science as a body of knowledge (NRC, 2007). All activities in science take place within an orienting framework of conceptual knowledge, connected not only to students’ prior knowledge but also to the more sophisticated understandings they are expected to develop as a result of instruction. Atkin (1966) argued that the process of science is embedded within the content it seeks to enrich and define (Rutherford, 1960).

The epistemic domain is based upon students knowing how scientific knowledge is generated. Evidence collected through students’ own scientific investigations is essential in this regard. Students should examine and evaluate the quality of evidence and then interpret evidence to develop explanations for phenomena (NRC, 1996, 2001, 2007). Then, a bridge should be provided to students so that they learn that their own process of collecting, evaluating, and interpreting evidence is similar to the practice of real scientists (e.g., Bell, Lederman, & Abd-El-Khalick, 1998). Furthermore, students should learn that scientific knowledge is subject to change in the face of new evidence or new interpretations of old evidence.

The social domain consists of the collaborative and communicative processes by which scientific knowledge is constructed, or the process by which students participate in scientific practices (NRC, 2007). Duschl (2003) described the social processes involved with communicating scientific ideas and understandings and emphasizes the importance of students making public their ideas through argument, modeling, and other modes of representation to help them learn to examine and evaluate their developing understanding of science. Since science is a collaborative enterprise, students should be encouraged to work in groups to reason collectively and reach decisions together.

To these three domains we add a fourth category, which we will call procedural, subdivided from Duschl’s epistemic domain. The procedural domain describes the methods or “heuristics of discovery” (Bruner, 1961, p. 30): that is, asking scientifically oriented questions, designing experiments, executing procedures, and creating data representations (NRC, 2001; Schwab, 1962). This distinction is important to make because inquiry-based teaching reforms are often described as...
hands-on activities where students manipulate materials and collect their own data, independent of engaging in the process of evaluating the data they collected.

Guidance Dimension of Inquiry
In addition to the cognitive demands placed on the student as defined earlier, another feature of reform-based science teaching is a constant series of transitions of responsibility for learning from the teacher to the student, and then back to the teacher. Rather than being passive recipients of scientific knowledge, students in inquiry-based lessons are actively engaged in constructing their own understandings. Guidance in science education is often characterized as varying with respect to what is left open to students to define or provided by the teacher or curriculum (Schwab, 1962; Shulman & Tamir, 1973). The NRC’s *Inquiry and the National Science Education Standards* defined a continuum of directedness across the “essential features” of inquiry, varying across the extent of learner self-direction and direction from the teacher or materials (NRC, 2001, p. 29).

Thus, the guidance dimension of inquiry distinguishes between reform conditions in which the teacher or student leads the activity. In this way, inquiry-based teaching can be thought of as part of a continuum of guidance. This continuum is bordered on one end by traditional *teacher-led* instruction in which the teacher tells students the answers they are expected to learn. At the other end of the continuum, students lead the activity, an approximation of discovery learning. We recognize a middle ground of *teacher-guided inquiry* in which the teacher actively guides students’ activities in a reform-oriented science lesson (Figure 1).

Method
The purpose of this review is to compare and contrast the effects that have been found for inquiry-based teaching as a function of the different dimensions we have defined earlier. The sections that follow will describe in detail our literature search methods and selection criteria, coding process, and meta-analytic procedure.

Literature Search Methods and Selection Process
This article was developed as an extension of a larger meta-analysis (Seidel & Shavelson, 2007) that explored the effects of different teaching interventions on student learning across disciplines. The sample of studies selected for the present meta-analysis began with the subset of studies from that previous report, which were conducted in science education, and extended that sample with additional search terms and years of publication.

We chose to focus upon the effects of inquiry-based teaching during the 10 years following publication of the *National Science Education Standards* (NRC, 1996), a time period during which the reform spotlight was intensely focused upon
this approach to science teaching reform (1996–2006). In this way, this meta-
analysis presents a historical analysis of the effects of a particular approach during
the time it was a popular “buzz word” in science education.

When adding to Seidel and Shavelson’s (2007) list of search terms, we consid-
ered that science educators and science education researchers use a rich vocabu-
lary beyond inquiry to describe the reforms they are developing and studying. To
name just a few, reforms in science teaching in the past 15 years have been called
inquiry-based teaching and learning (e.g., NRC, 2001), authentic inquiry (Chinn
& Malhotra, 2002), model-based inquiry (Windschitl, Thompson, & Braaten,
2008), modeling and argumentation (McNeill, Lizotte, Krajcik, & Marx, 2006),
project-based science (Singer, Marx, Krajcik, & Chambers, 2000), hands-on sci-
ence (Pine et al., 2006), and constructivist science (Hardy, Jonen, Möller, & Stern,
2006). Klahr and Li (2005) argued that this profusion of terminology has made
difficult the advancement of research in science education because of the lack of
clear operational terms for what does and does not constitute a reform. In fact,
Hmelo-Silver et al. (2007) argued that critiques of these approaches to science
teaching (e.g., Kirschner et al., 2006) indicate a misunderstanding as to what con-
stitutes each kind of reform.

To capture the diversity of terms used to describe inquiry-based reforms, we
appended the keywords from Seidel and Shavelson’s (2007) meta-analysis and
conducted a new search in Web of Science and ERIC with the following teaching
keywords: effective instruction, instructional effectiveness, direct instruction,
teacher effectiveness, mastery learning, constructivist teaching, science instruc-
tion, classrooms, science teaching, and inquiry. We crossed each of these key-
words with the output keywords achievement, competencies, interest, motivation,
engagement, and attainment. We did not use the term learning alone because it
yielded too many different kinds of studies.

In order to focus upon the effects of science teaching reforms on student learn-
ing in mainstream K–12 classroom settings, we excluded studies that did not have
experimental or quasi-experimental designs and that focused specifically on stu-
dents with disabilities, special education, or interventions that took place outside
regular K–12 science classrooms. Our search terms yielded an initial sample of
1,625 studies. We iteratively reduced this initial sample to smaller and smaller
pools of studies by reading abstracts and, when necessary, full papers, eliminating
studies that did not satisfy our inclusion criteria. This process yielded a reduced
sample of 59 papers. We then coded each paper according to features of its exper-
imental design and outcome measure, retaining only papers published in English;
those that had a pre-post, two-group design; and those that used a cognitive out-
come measure with data provided to calculate an effect size. This left a total of 15
papers. Five of these studies did not include data necessary for calculating effect
sizes. After contacting these authors, we were able to obtain necessary data for 4
papers. The first search thus yielded 14 papers for inclusion.

However, this initial search did not yield important experimental studies known
to the first author, such as Chang and Mao’s (1999) comparison of inquiry-group
versus traditional instruction and Hardy et al.’s (2006) study of constructivist ver-
sus low instructional support. We therefore cross-referenced these articles with the
original search terms and conducted an additional search with new terms. We

307

Downloaded from http://rer.aera.net at UNIV OF ROCHESTER LIBRARY on December 4, 2012
crossed support, traditional, and instruction with the reduced outcome search terms understanding, achievement, attainment, and effect. After reading abstracts and eliminating studies based on our inclusion criteria, this secondary search yielded 15 additional papers, 7 of which had complete data. After emailing authors we received data from 1 additional paper. Thus, in the end, our two searches yielded a total of 22 papers for inclusion in our meta-analysis (see Figure 2 for a schematic representation of the selection process).

As we describe earlier, we attempted to retrieve as many published studies as possible that met our search criteria. Despite the fact that our review of the literature was exhaustive, we are aware that we still may have missed some relevant publications. However, we believe that the search was thorough enough to present a collection of publications that is representative of the field between the years 1996 and 2006.

**Multiple Contrasts**

Multiple contrasts were reported in several of the papers in our final sample. To preserve these individual studies within papers, we included contrasts between different permutations of the treatment (e.g., teacher led vs. student led), contrasts that separated out the effects of the treatment in different subject areas, and contrasts that showed the effect of the treatment based upon multiple outcome measures. Three papers (Hardy et al., 2006; Huffman, 1997; Pedaste & Sarapuu, 2006) disaggregated outcome measures by specific item or content-specific groups of items. In these situations we reported only an aggregate for the outcome measure as a whole. These criteria yielded 37 studies from our sample of 22 papers.
Classification Process

A central question of this meta-analysis was whether the use of the cognitive and guidance dimensions of inquiry would help us better identify the instructional approaches that were contrasted in each study. As we described previously, each study contrasted different elements of these dimensions; thus, we coded the experimental and control condition for each study according to the cognitive and guidance dimensions to identify which dimensions were being explicitly contrasted.

Design and Structure Codes

Two coders used a process of consensus coding to code studies according to the research design, the way the studies self-identified their treatments (e.g., inquiry, discovery), the grade level of the study, and the country in which the study took place.

Inquiry Codes

The four inquiry domains were operationalized into subcategories, which were then used to code the description of the teaching in each condition. In this way, we were able to identify the contrast between the control and treatment groups (e.g., conceptual only control vs. full model of inquiry treatment). The inquiry codes are shown in Table 2. Two raters independently applied the inquiry codes to the experimental groups in 11 of the 22 papers (representing 19 of 37 studies, roughly half of the total sample). Raters used each code dichotomously to indicate the presence or absence of each of the four inquiry domains for a total of 44 codes given by each rater. The two raters applied the same code 33 out of 44 times, establishing 75% direct interrater agreement. Disagreements were resolved through adjudication, and one rater then proceeded to code the remainder of the studies.

### Table 2

<table>
<thead>
<tr>
<th>Domain of inquiry</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedural</td>
<td>Asking scientifically oriented questions</td>
</tr>
<tr>
<td></td>
<td>Experimental design</td>
</tr>
<tr>
<td></td>
<td>Executing scientific procedures</td>
</tr>
<tr>
<td></td>
<td>Recording data</td>
</tr>
<tr>
<td></td>
<td>Representing data</td>
</tr>
<tr>
<td></td>
<td>Hands-on</td>
</tr>
<tr>
<td>Epistemic</td>
<td>Nature of science</td>
</tr>
<tr>
<td></td>
<td>Drawing conclusions based on evidence</td>
</tr>
<tr>
<td></td>
<td>Generating and revising theories</td>
</tr>
<tr>
<td>Conceptual</td>
<td>Drawing on/connecting to prior knowledge</td>
</tr>
<tr>
<td></td>
<td>Eliciting students’ ideas/mental models</td>
</tr>
<tr>
<td></td>
<td>Providing conceptually oriented feedback</td>
</tr>
<tr>
<td>Social</td>
<td>Participating in class discussions</td>
</tr>
<tr>
<td></td>
<td>Arguing/debating scientific ideas</td>
</tr>
<tr>
<td></td>
<td>Presentations</td>
</tr>
<tr>
<td></td>
<td>Working collaboratively</td>
</tr>
</tbody>
</table>

A central question of this meta-analysis was whether the use of the cognitive and guidance dimensions of inquiry would help us better identify the instructional approaches that were contrasted in each study. As we described previously, each study contrasted different elements of these dimensions; thus, we coded the experimental and control condition for each study according to the cognitive and guidance dimensions to identify which dimensions were being explicitly contrasted.

Design and Structure Codes

Two coders used a process of consensus coding to code studies according to the research design, the way the studies self-identified their treatments (e.g., inquiry, discovery), the grade level of the study, and the country in which the study took place.

Inquiry Codes

The four inquiry domains were operationalized into subcategories, which were then used to code the description of the teaching in each condition. In this way, we were able to identify the contrast between the control and treatment groups (e.g., conceptual only control vs. full model of inquiry treatment). The inquiry codes are shown in Table 2. Two raters independently applied the inquiry codes to the experimental groups in 11 of the 22 papers (representing 19 of 37 studies, roughly half of the total sample). Raters used each code dichotomously to indicate the presence or absence of each of the four inquiry domains for a total of 44 codes given by each rater. The two raters applied the same code 33 out of 44 times, establishing 75% direct interrater agreement. Disagreements were resolved through adjudication, and one rater then proceeded to code the remainder of the studies.
Guidance Codes
In coding the studies, we identified four different kinds of guidance contrasts: traditional instruction versus student-led reform, traditional instruction versus teacher-led reform, student- versus teacher-led, and guidance dimension undefined. Traditional comparison groups were always treated as controls. As a rule of thumb, we labeled the student-led condition as the control in studies that contrasted student-led conditions with teacher-led conditions. We made this distinction following literature emphasizing the importance of teacher guidance in inquiry-based science teaching reforms (e.g., Hmelo-Silver et al., 2007). Studies falling into the undefined group contrasted some kind of visual representation (e.g., Kim, 2006) with another kind of representation or one computer-based intervention with another (e.g., White & Frederiksen, 1998), and thus guidance was not explicitly a feature of the study’s design. In some studies, we coded the contrasts differently than the authors had identified them to ensure consistency across studies within our own sample. For example, Gläser-Zikuda, Fuß, Laukenmann, Metz, and Randler (2005) identified their treatment as student centered, but the intervention was a phase of student-centered instruction followed by teacher-centered time focusing on correction of mistakes and assessment. Thus, we coded this study as teacher-guided reform. Two raters applied the guidance codes through a process of consensus coding.

Computing Effect Sizes
The effect size statistic allows for the comparison of treatment effects across studies where the dependent variable may not necessarily be operationalized in the same way (Lipsey & Wilson, 2001, p. 48). A commonly used effect size formula is Glass’s Δ (Glass, 1976), represented as

\[
ES = \frac{(\bar{X}_{\text{Treatment}} - \bar{X}_{\text{Control}})}{SD_{\text{Control}}},
\]

where \( \bar{X} \) represents the average scores for the treatment and control groups and \( SD_{\text{Control}} \) is the standard deviation of the control group.

This formula can be altered slightly when different experimental designs are employed (see Lipsey & Wilson, 2001), such as in situations where the study involves pre- and posttest measures for both a treatment and control group (Becker, 1988; Morris, 2008). In these situations the effect size may be different from that calculated in the standard approach (i.e., Glass’s Δ) if the mean pretest score and standard deviations across the treatment and control groups are taken into account. This is particularly important in our meta-analysis because over half of the studies (23 of 37) had higher pretest scores for the treatment group than the control group. In order to mitigate the threat of selection bias in these data, pretest results will be included in the effect size calculation. Morris (2008) conducted an analysis of several different methods of calculating effect size for studies with pretest-posttest-control designs and found the following equation to yield more precise effect size estimates as a result of a smaller sampling variance when compared with other approaches.
Meta-Analysis of Inquiry-Based Teaching

\[ ES_{\text{Pre-/Post-Test Two Groups}} = \frac{(\bar{X}_{T\text{-Post}} - \bar{X}_{T\text{-Pre}}) - (\bar{X}_{C\text{-Post}} - \bar{X}_{C\text{-Pre}})}{SD_{\text{Pooled_Pre}}}, \]  

(2)

In this equation, \( \bar{X} \) represents a test score mean for treatment and control conditions (subscripts “T” and “C”) administered at the beginning and end of a study period (subscripts “PRE” and “POST”), and SD is computed as the weighted average of the standard deviations across treatment and control groups.

However, Morris’ (2008) equation is not directly applicable to our data as his analysis was conducted in the domain of organizational interventions. In the case of achievement data, one would expect variance to increase from pretest to post-test, and thus using the pooled pretest SD as the denominator would lead to inflated effect sizes. Indeed, Morris’ approach yielded a larger effect size with greater variance as compared to using \( SD_{\text{Control_Post}} \) in the denominator.\(^1\) Therefore, we used a modified version of Morris’ equation as follows:

\[ ES_{\text{Pre-/Post-Test Two Groups}} = \frac{(\bar{X}_{T\text{-Post}} - \bar{X}_{T\text{-Pre}}) - (\bar{X}_{C\text{-Post}} - \bar{X}_{C\text{-Pre}})}{SD_{\text{C\text{-Post}}}}, \]  

(3)

Since all the studies in our meta-analysis—even those with random assignment—involve a pre-post, two-group design, and since the majority of posttest SDs were greater than pretest SDs, we calculated mean effect sizes using Equation 3. Use of \( SD_{\text{Control\text{-Post}}} \) is also consistent with the denominator used in the most recent meta-analysis of student achievement in science (Schroeder et al. 2007).

Although many meta-analysts often weight study effect sizes based on the number of students in the study, we have chosen not to do so for two reasons based on critiques of weighting by sample size. First, the concept of weighting by sample size stems from the assumption that each study can be considered a replication of the other and that each study is like a random draw from some implicit population of studies, with each sample of students a random draw from some implicit population of students; however, this assumption falls short since research studies take place at different times and at schools with vastly differing student populations and curricula. Second, weighting based on sample size is most relevant when the focus of a meta-analysis is to produce a single number that represents the main effect of a given intervention. The principal focus in this study is to examine the variability in effect sizes by the way that inquiry-based teaching has been defined.

**Results**

*Descriptive Information*

The studies represented a range of subject matters, were performed in a variety of countries, and examined teaching in elementary, middle, and secondary school. Descriptive codes are listed in Table 3 for all 37 studies from the 22 papers. Of the 22 papers included in our analysis, 5 presented studies conducted in only the
**TABLE 3**
List of included studies, codes applied, and effect sizes

<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Study number</th>
<th>Country</th>
<th>Subject matter</th>
<th>Level</th>
<th>Treatment</th>
<th>Control</th>
<th>Inquiry contrast</th>
<th>Guid-ance</th>
<th>Total n</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexander, Fives, Buehl, and Mulhern (2002)</td>
<td>1</td>
<td>United States</td>
<td>Astronomy</td>
<td>Middle</td>
<td>PECS</td>
<td>C</td>
<td>PES</td>
<td>1</td>
<td>110</td>
<td>1.737</td>
</tr>
<tr>
<td>Alexander et al. (2002)</td>
<td>2</td>
<td>United States</td>
<td>Astronomy</td>
<td>Middle</td>
<td>PECS</td>
<td>C</td>
<td>PES</td>
<td>0</td>
<td>126</td>
<td>0.956</td>
</tr>
<tr>
<td>Ardac and Akaygun (2004)</td>
<td>1</td>
<td>Turkey</td>
<td>Chemistry</td>
<td>Middle</td>
<td>C</td>
<td>C</td>
<td>None</td>
<td>3</td>
<td>49</td>
<td>0.622</td>
</tr>
<tr>
<td>Balci, Cakiroglu, and Tekkaya (2006)</td>
<td>1</td>
<td>Turkey</td>
<td>Biology</td>
<td>Elementary</td>
<td>PECS</td>
<td>C</td>
<td>PES</td>
<td>1</td>
<td>67</td>
<td>0.779</td>
</tr>
<tr>
<td>Balci et al. (2006)</td>
<td>2</td>
<td>Turkey</td>
<td>Biology</td>
<td>Elementary</td>
<td>EC</td>
<td>C</td>
<td>E</td>
<td>1</td>
<td>68</td>
<td>0.545</td>
</tr>
<tr>
<td>Cepni (2006)</td>
<td>1</td>
<td>Turkey</td>
<td>Biology</td>
<td>High</td>
<td>C</td>
<td>C</td>
<td>None</td>
<td>3</td>
<td>52</td>
<td>1.214</td>
</tr>
<tr>
<td>Chang (2002)</td>
<td>1</td>
<td>Taiwan</td>
<td>Earth science</td>
<td>High</td>
<td>PEC</td>
<td>PEC</td>
<td>None</td>
<td>2</td>
<td>244</td>
<td>0.043</td>
</tr>
<tr>
<td>Chang and Mao (1999)</td>
<td>1</td>
<td>Taiwan</td>
<td>Earth science</td>
<td>Middle</td>
<td>PEC</td>
<td>C</td>
<td>PES</td>
<td>0</td>
<td>612</td>
<td>0.148</td>
</tr>
<tr>
<td>Chang and Tsai (2005)</td>
<td>1</td>
<td>Taiwan</td>
<td>Earth science</td>
<td>High</td>
<td>PEC</td>
<td>PEC</td>
<td>None</td>
<td>2</td>
<td>347</td>
<td>0.007</td>
</tr>
<tr>
<td>Chang, Hsiao, and Barufaldi (2006)</td>
<td>1</td>
<td>Taiwan</td>
<td>Earth science</td>
<td>High</td>
<td>PECS</td>
<td>C</td>
<td>PES</td>
<td>1</td>
<td>155</td>
<td>0.049</td>
</tr>
<tr>
<td>Diakidoy and Kendecou (2001)</td>
<td>1</td>
<td>Cyprus</td>
<td>Astronomy</td>
<td>Elementary</td>
<td>ECS</td>
<td>C</td>
<td>ES</td>
<td>1</td>
<td>54</td>
<td>1.369</td>
</tr>
<tr>
<td>Ertepinar and Geban (1996)</td>
<td>1</td>
<td>Turkey</td>
<td>Physical science/intro chemistry</td>
<td>Secondary</td>
<td>PECS</td>
<td>C</td>
<td>PES</td>
<td>1</td>
<td>43</td>
<td>0.652</td>
</tr>
<tr>
<td>Gläser-Zikuda (2005)</td>
<td>1</td>
<td>Germany</td>
<td>Physics</td>
<td>Grades 8–9</td>
<td>CS</td>
<td>C</td>
<td>S</td>
<td>1</td>
<td>378</td>
<td>0.185</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Study number</td>
<td>Country</td>
<td>Subject matter</td>
<td>Level</td>
<td>Treatment</td>
<td>Control</td>
<td>Inquiry contrast</td>
<td>Guid-ance</td>
<td>Total n</td>
<td>Effect size</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>--------------</td>
<td>------------------</td>
<td>----------------------</td>
<td>-------------</td>
<td>-----------</td>
<td>---------</td>
<td>------------------</td>
<td>-----------</td>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>Gläser-Zikuda (2005)</td>
<td>2</td>
<td>Germany</td>
<td>Biology</td>
<td>Grades 8–9</td>
<td>CS</td>
<td>C</td>
<td>S</td>
<td>1</td>
<td>281</td>
<td>−0.005</td>
</tr>
<tr>
<td>Hardy et al. (2006)</td>
<td>1</td>
<td>Germany</td>
<td>Physical science</td>
<td>Elementary</td>
<td>PECS</td>
<td>0</td>
<td>PECS</td>
<td>0</td>
<td>95</td>
<td>0.236</td>
</tr>
<tr>
<td>Hardy et al. (2006)</td>
<td>2</td>
<td>Germany</td>
<td>Physical science</td>
<td>Elementary</td>
<td>PECS</td>
<td>0</td>
<td>PECS</td>
<td>1</td>
<td>102</td>
<td>0.251</td>
</tr>
<tr>
<td>Hardy et al. (2006)</td>
<td>3</td>
<td>Germany</td>
<td>Physical science</td>
<td>Elementary</td>
<td>PECS</td>
<td>PECS</td>
<td>None</td>
<td>2</td>
<td>125</td>
<td>0.013</td>
</tr>
<tr>
<td>Huffman (1997)</td>
<td>1</td>
<td>United States</td>
<td>Physics</td>
<td>High</td>
<td>CS</td>
<td>C</td>
<td>S</td>
<td>3</td>
<td>145</td>
<td>1.045</td>
</tr>
<tr>
<td>Huffman (1997)</td>
<td>2</td>
<td>United States</td>
<td>Physics</td>
<td>High</td>
<td>CS</td>
<td>C</td>
<td>S</td>
<td>3</td>
<td>145</td>
<td>0.206</td>
</tr>
<tr>
<td>Huffman (1997)</td>
<td>3</td>
<td>United States</td>
<td>Physics</td>
<td>High</td>
<td>CS</td>
<td>C</td>
<td>S</td>
<td>3</td>
<td>145</td>
<td>−0.240</td>
</tr>
<tr>
<td>Huffman (1997)</td>
<td>4</td>
<td>United States</td>
<td>Physics</td>
<td>High</td>
<td>CS</td>
<td>C</td>
<td>S</td>
<td>3</td>
<td>145</td>
<td>−0.168</td>
</tr>
<tr>
<td>Keselman (2003)</td>
<td>1</td>
<td>United States</td>
<td>Multivariable causality</td>
<td>Middle</td>
<td>PECS</td>
<td>PS</td>
<td>EC</td>
<td>3</td>
<td>46</td>
<td>0.625</td>
</tr>
<tr>
<td>Keselman (2003)</td>
<td>2</td>
<td>United States</td>
<td>Multivariable causality</td>
<td>Middle</td>
<td>PES</td>
<td>PS</td>
<td>E</td>
<td>3</td>
<td>46</td>
<td>0.786</td>
</tr>
<tr>
<td>Kiboss (2005)</td>
<td>1</td>
<td>Kenya</td>
<td>Physical science</td>
<td>Secondary</td>
<td>C</td>
<td>C</td>
<td>None</td>
<td>3</td>
<td>79</td>
<td>1.547</td>
</tr>
<tr>
<td>Kiboss (2005)</td>
<td>2</td>
<td>Kenya</td>
<td>Physical science</td>
<td>Secondary</td>
<td>C</td>
<td>C</td>
<td>None</td>
<td>3</td>
<td>79</td>
<td>0.929</td>
</tr>
<tr>
<td>Kim (2006)</td>
<td>1</td>
<td>United States</td>
<td>Earth science</td>
<td>Elementary</td>
<td>PC</td>
<td>PC</td>
<td>None</td>
<td>3</td>
<td>41</td>
<td>0.472</td>
</tr>
<tr>
<td>Lazarowitz, Baird, Bowlden, and</td>
<td>1</td>
<td>Israel/United</td>
<td>Biology</td>
<td>High</td>
<td>PCS</td>
<td>PC</td>
<td>S</td>
<td>0</td>
<td>113</td>
<td>0.187</td>
</tr>
<tr>
<td>Hertz-Lazarowitz (1996)</td>
<td></td>
<td>States</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Author(s)</th>
<th>Study number</th>
<th>Country</th>
<th>Subject matter</th>
<th>Level</th>
<th>Treatment</th>
<th>Control</th>
<th>Inquiry contrast</th>
<th>Guid-ance</th>
<th>Total n</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lazarowitz et al. (1996)</td>
<td>2</td>
<td>Israel/United States</td>
<td>Biology</td>
<td>High</td>
<td>PCS</td>
<td>PC</td>
<td>S</td>
<td>0</td>
<td>113</td>
<td>−0.300</td>
</tr>
<tr>
<td>Pedaste and Sarapuu (2006)</td>
<td>1</td>
<td>Estonia</td>
<td>Environmental science and ecology</td>
<td>Grades 6–12</td>
<td>PEC</td>
<td>PEC</td>
<td>None</td>
<td>3</td>
<td>125</td>
<td>0.393</td>
</tr>
<tr>
<td>Reid et al. (2003)</td>
<td>1</td>
<td>United Kingdom</td>
<td>Physics</td>
<td>Middle</td>
<td>PEC</td>
<td>P</td>
<td>EC</td>
<td>2</td>
<td>38</td>
<td>−0.040</td>
</tr>
<tr>
<td>Reid et al. (2003)</td>
<td>2</td>
<td>United Kingdom</td>
<td>Physics</td>
<td>Middle</td>
<td>PEC</td>
<td>P</td>
<td>EC</td>
<td>2</td>
<td>38</td>
<td>−0.012</td>
</tr>
<tr>
<td>Reid et al. (2003)</td>
<td>3</td>
<td>United Kingdom</td>
<td>Physics</td>
<td>Middle</td>
<td>PC</td>
<td>P</td>
<td>C</td>
<td>2</td>
<td>38</td>
<td>0.021</td>
</tr>
<tr>
<td>Ugwu and Soyibo (2004)</td>
<td>1</td>
<td>Jamaica</td>
<td>Biology</td>
<td>Middle</td>
<td>CS</td>
<td>CS</td>
<td>None</td>
<td>3</td>
<td>266</td>
<td>1.190</td>
</tr>
<tr>
<td>Ugwu and Soyibo (2004)</td>
<td>2</td>
<td>Jamaica</td>
<td>Biology</td>
<td>Middle</td>
<td>C</td>
<td>C</td>
<td>None</td>
<td>3</td>
<td>351</td>
<td>0.705</td>
</tr>
<tr>
<td>Ugwu and Soyibo (2004)</td>
<td>3</td>
<td>Jamaica</td>
<td>Biology</td>
<td>Middle</td>
<td>C</td>
<td>C</td>
<td>None</td>
<td>3</td>
<td>315</td>
<td>1.617</td>
</tr>
<tr>
<td>Uzuntiryaki and Geban (2005)</td>
<td>1</td>
<td>Turkey</td>
<td>Chemistry</td>
<td>Middle</td>
<td>EC</td>
<td>C</td>
<td>E</td>
<td>1</td>
<td>64</td>
<td>0.920</td>
</tr>
<tr>
<td>White and Frederiksen (1998)</td>
<td>1</td>
<td>United States</td>
<td>Physics</td>
<td>High</td>
<td>PECS</td>
<td>PECS</td>
<td>None</td>
<td>3</td>
<td>302</td>
<td>−0.064</td>
</tr>
</tbody>
</table>

*Note.* P = procedural; E = epistemic; C = conceptual; S = social; 0 = traditional versus student led; 1 = traditional versus teacher led; 2 = student centered versus teacher centered; 3 = guidance not contrasted.
United States, 5 in Turkey, 4 in Taiwan, and 2 in Germany, with additional studies from Cyprus, Estonia, Israel/United States, Jamaica, Kenya, and the United Kingdom. This international sample of studies spanned a range of subject areas, with 4 taking place in biology, 8 in Earth and space science, 6 in physics/physical science, 3 in chemistry, and 1 that studied causality without a specific content focus. In terms of school type, 4 studies took place in elementary school, 7 in middle school, 7 in high school, and 4 in what was termed secondary school spanning Grades 6 to 12.

The number of papers published each year between 1996 and 2006 varied. Two studies were published in 1996; 4 papers were published each year in 1997, 1998, 1999, and 2001; 6 papers were published between 2002 and 2004; and finally, 10 papers were published between 2005 and 2006. This suggests a possible upward trend in experimental studies in response to growing governmental initiatives in the United States in the early 2000s (e.g., Shavelson & Towne, 2002).

Only three papers (Cepni, Tas, & Köse, 2006; Kiboss & Ogunniyi, 2005; Reid, Zhang, & Chen, 2003) involved the random assignment of students to conditions, or true experimental studies. A second group of papers \( n = 13 \) involved random assignment of whole classes to conditions. Four papers involved nonrandom assignment of whole classes to conditions, or quasi-experimental studies, and the remaining two studies provided no information about assignment.

Only four of the papers included in this analysis self-identified as studying what they called inquiry-based teaching. The other papers used a variety of terms to describe the treatments, including cooperative learning, problem solving, discovery learning, instructional support, investigative approach, and constructivist learning. A subset of papers focused on computer-based interventions described as simulations and computer-assisted or multimedia approaches. The control groups in each study were described with a different set of terms: textbook approach, traditional instruction, direct instruction, individual mastery learning, regular instruction, baseline, and practice alone.

**Overall Mean Effect Size**

The distribution of individual effect sizes for all 37 studies included is shown in Figure 3. The overall mean effect size was .50 \( (SD = 0.56) \). This mean is greater than all of the mean effect sizes reported by the previous meta-analyses reviewed at the beginning of this article (see Table 2) with the exception of Schroeder et al. (2007), who found an effect of .65.

As Figure 3 indicates, although 7 studies had negative effects, 23 studies had effect sizes between 0 and 1, and 7 studies had an effect size between 1 and 2. When we explored the relationship between effect size and sample size, we found that the studies with larger sample sizes had effect sizes within the same range of the other studies with smaller sample sizes (Figure 4).

**Effect Size Contrasts on the Cognitive Dimension**

The first step in our analysis was to compare effect sizes using the cognitive dimension of inquiry-based teaching. The studies included in our meta-analysis used multiple reform terms to describe their treatment groups and multiple terms to describe their control groups. In this section, we present results of the meta-analysis according to the domain of inquiry being contrasted. As shorthand, we
will refer to each combination of domains by abbreviations. For example, a study coded for all four domains of the cognitive dimension—procedural, epistemic, conceptual, and social domains—will be referred to as PECS, and a study with only the procedural and conceptual domains will be referred to as PC. We will also provide examples from the studies that follow.

The most common code given to the treatment group was PECS; 10 of the 37 studies included this kind of treatment in its analysis. The next most commonly coded treatment conditions were CS (7 studies), C (6 studies), and PEC (5 studies). In total, there were 10 different combinations of the four domains represented in the treatment groups in the sample of 37 studies. With respect to the control group, 21 studies were only coded for the conceptual domain. In total, there were eight different combinations of the domains in the control groups. These results indicate that there were a variety of treatments and control groups represented in the sample of studies.

FIGURE 3. Distribution of effect sizes (N = 37 studies).

Note. Bin width equals one standard deviation.
We do not argue that the four cognitive inquiry domains were independent of each other; thus, we did not interpret effect sizes by inquiry domains included in the treatment. Rather, we explored the difference between the domain codes given to the treatment and control groups in each contrast. This difference tells us what elements of the intervention were being varied in the study, and thus to which domain of inquiry the effect of the intervention might be attributed. For example, the conditions contrasted in Chang and Mao (1999) were coded as PECS versus C, or a contrast of PES. In the experimental condition, students worked in teams to generate data (procedural), present and discuss data (social), and construct explanations on the basis of that data (epistemic), all embedded within the content of the earth-moon-sun system (conceptual). In the control condition, students experienced lectures based on the same concepts with occasional demonstrations for key ideas (conceptual). In the case of Diakidoy and Kendeou (2001), students in the experimental condition began by discussing their prior ideas about the shape of the earth and gravity and moved on to focusing on the concepts of speed of rotation.
and perception of movement (conceptual), discussing the results of a demonstration (social) and formulating hypotheses and explanations about how day changes into night (epistemic). In the control condition, students read booklets that focused on the same concepts, where the teacher described and defined each concept, wrote those concepts on the board, and had students answer questions from the booklet orally and in writing (conceptual).

Surprisingly, for 13 of the studies, there was no difference in the inquiry domain codes in the treatment and control groups, and thus there was no contrast in the cognitive dimension of inquiry to report. For several of the studies (Ardac & Akaygun, 2004; Cepni et al., 2006; Chang, 2002; Chang & Tsai, 2005; Kiboss & Ogunniyi, 2005; Pedaste & Sarapuu, 2006), the contrast did not show up in the cognitive dimension codes because the contrast was between different types of computer-assisted environments or the same curriculum when it was given through regular teaching or led by a computer. However, two other studies (Hardy et al., 2006; White & Frederiksen, 1998) were coded as having PECS present in both the control and experimental groups and thus were contrasting features other than the cognitive domains of inquiry as defined for this analysis. In Hardy et al. (2006), the contrast between the conditions was based on the extent to which the student activities were guided by the teacher or students, and this contrast is captured with the guidance codes. In the case of White and Frederiksen (1998), the contrast between the conditions was in the extent to which metacognition was integrated into students’ activities, a variation that is subsumed within our definition of the conceptual domain. However, our coding system did not capture this variation because the control condition in this study was also embedded in conceptual content.

For the remaining 22 studies, we were able to examine contrasts in treatment and control conditions as a function of our PECS codes. Table 4 shows that the 3 studies that explicitly contrasted the epistemic domain of inquiry had the largest mean effect size on student learning, followed by the 6 studies that contrasted the procedural, epistemic, and social domains. The most common contrast was social, which had the lowest effect size from the five contrasts presented previously.

### Effect Size Contrasts on the Guidance Dimension

The next step in our analysis was to compare effect sizes using the guidance dimension of inquiry-based teaching. As described in the methods section, we

---

**TABLE 4**

Mean effect size by model of inquiry contrasted

<table>
<thead>
<tr>
<th>Contrast</th>
<th>N studies</th>
<th>N papers</th>
<th>Min</th>
<th>Max</th>
<th>SD</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC</td>
<td>3</td>
<td>2</td>
<td>-.04</td>
<td>0.63</td>
<td>.38</td>
<td>.19</td>
<td>-.01</td>
</tr>
<tr>
<td>S</td>
<td>8</td>
<td>3</td>
<td>-.30</td>
<td>1.05</td>
<td>.43</td>
<td>.11</td>
<td>.09</td>
</tr>
<tr>
<td>PECS</td>
<td>2</td>
<td>1</td>
<td>.24</td>
<td>.25</td>
<td>.01</td>
<td>.24</td>
<td>.24</td>
</tr>
<tr>
<td>PES</td>
<td>6</td>
<td>5</td>
<td>.05</td>
<td>1.74</td>
<td>.61</td>
<td>.72</td>
<td>.72</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>3</td>
<td>.55</td>
<td>.92</td>
<td>.19</td>
<td>.75</td>
<td>.79</td>
</tr>
</tbody>
</table>

*Note: Overall mean effect size = .50 across the 37 studies. Table does not provide mean effect size for studies that did not explicitly study guidance or for which there was only one study in a category. P = procedural; E = epistemic; C = conceptual; S = social.*
identified four categories of guidance within the set of studies included in the meta-analysis: teacher-led reform, student-led reform, student versus teacher led, and other. A total of 16 studies that did not explicitly contrast guidance between the treatment and control groups were coded as guidance undefined.

The first three categories were applied as contrasts in the 21 studies shown in Table 5. Table 5 indicates that the 10 studies that explicitly contrasted teacher-guided reform versus traditional conditions had a higher mean effect size than the 5 that contrasted student-led reform versus traditional. The 6 studies that contrasted student- versus teacher-led reform had a mean effect size of .01. Finally, we compared the studies by guidance and inquiry domain contrast (Table 6). For the 10 studies that we coded as guidance undefined and that had no inquiry domain contrast, the mean effect size was large, suggesting that the computer-based interventions and representations explored in these studies successfully support students in learning science. We found the largest effect sizes for studies that contrasted traditional versus teacher-led reform and that varied PES between groups or E between groups (Figure 5).

### TABLE 5
**Effect sizes by guidance contrasted in study**

<table>
<thead>
<tr>
<th>Guidance</th>
<th>N studies</th>
<th>Min</th>
<th>Max</th>
<th>SD</th>
<th>Mean</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student led versus teacher led</td>
<td>6</td>
<td>-.04</td>
<td>.04</td>
<td>.03</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>Traditional versus student-led reform</td>
<td>5</td>
<td>-.30</td>
<td>.96</td>
<td>.45</td>
<td>.25</td>
<td>.19</td>
</tr>
<tr>
<td>Traditional versus teacher-led reform</td>
<td>10</td>
<td>-.01</td>
<td>1.74</td>
<td>.57</td>
<td>.65</td>
<td>.60</td>
</tr>
</tbody>
</table>

*Note. Overall mean effect size = .50 across the 37 studies. Table does not provide mean effect size for studies that did not explicitly study guidance or for which there was only one study in a category.*

### TABLE 6
**Median effect size by guidance and inquiry domain contrast**

<table>
<thead>
<tr>
<th>Inquiry domain contrast</th>
<th>Guidance</th>
<th>None</th>
<th>EC</th>
<th>S</th>
<th>PECS</th>
<th>PES</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional versus student-led reform</td>
<td>n = 2</td>
<td>.09</td>
<td></td>
<td></td>
<td>.80</td>
<td></td>
<td>.73</td>
</tr>
<tr>
<td>Traditional versus student-led reform</td>
<td>n = 2</td>
<td></td>
<td></td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student led versus teacher led</td>
<td>n = 3</td>
<td>.02</td>
<td></td>
<td></td>
<td>.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guidance undefined</td>
<td>n = 10</td>
<td></td>
<td></td>
<td></td>
<td>.86</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion

The purpose of this review was to compare and contrast the effects that have been found for studies of inquiry-based teaching published in the 10 years following release of the *National Science Education Standards* (NRC, 1996). We separated inquiry-based science teaching into cognitive and guidance dimensions and used this framework to arrive at a more nuanced understanding of the construct of inquiry-based science teaching for the sample of studies identified for inclusion.

The 37 studies that met our inclusion criteria had a mean effect size of .50. This mean was higher than any of the means reported in meta-analyses conducted in the 1980s (which ranged from .06 to .43) and approaches the .65 mean effect size found by Schroeder et al. (2007), even though there is no overlap in the papers included in the two studies. We attribute this lack of overlap to differences in search terms and inclusion criteria. Schroeder et al. used a different set of search...

FIGURE 5. *Location of effect sizes for guidance contrasts.*
terms that focused more on study design and content area. In addition, Schroeder et al.’s inclusion criteria differed in two significant ways. First, Schroeder et al. had a wider range of search years (1980–2004) and included only studies performed in the United States. Since only one of the U.S. studies in our sample fell outside Schroeder et al.’s search years (Kim, 2006), the remaining lack of overlap in studies must be attributed to differences in search techniques and terms.

In the following sections, we begin by discussing limitations of our study sample, the need for data standards for publication, and construct validity and limitations of the meta-analytic approach. We will then summarize and discuss our findings and their implications for science teaching and learning. Finally, we will close with implications for reform science teaching terminology.

Limitations of Study Sample

Our sample was drawn exclusively from papers published in peer-reviewed, research-oriented journals. We made this decision to focus on papers that had met the levels of rigor necessitated by the peer review process. Since these journals often do not publish studies with negative results, one might argue that we missed out on a larger range of findings available in unpublished reports and theses. Despite excluding these reports, we still found several studies within our sample that did present negative results. Furthermore, our mean effect size was within the range of those found in previous meta-analyses with different selection criteria, suggesting that our sample of studies, while limited, was still roughly representative of the field.

Unfortunately, we found very few experimental and quasi-experimental studies performed in K–12 science classrooms during the years sampled. In addition, although the field of science education scarcely needs a reminder, our study reminds us yet again that reform-oriented science teaching practices are difficult to describe, difficult to enact, and even more difficult to characterize.

Implications for Meta-Analytic Approach

Need for Data Standards for Publication

An unexpected and disheartening finding of this analysis was the number of studies that met our design criteria for selection but did not present the necessary data for calculating effect sizes. Many of these studies presented group means with standard errors or skipped presentation of means altogether in favor of $F$, $p$, $\sigma$, and $\chi^2$ values. We contacted the authors of 13 studies, and of the nine authors who responded, three could not send their data. Of those three authors, one could no longer access the original data, and one had lost the data. Although we acknowledge that these data may not always contribute to the substantive story being told through the answering of the research questions posed by original studies, and that strict word limits may force authors to omit data that they view as peripheral to their argument, omission of these data nevertheless makes it impossible to advance our understanding through the use of meta-analytic techniques. We argue that in order for the field to move forward, studies with experimental designs should always include basic relevant statistics such as means and standard deviations so that they may be included in future meta-analyses.
Construct Validity and Limitations of the Meta-Analytic Approach

A meta-analysis represents an attempt to synthesize causal inferences across studies, and each study will vary with respect to its context. Although a meta-analysis can facilitate causal generalization, these generalizations remain constrained by the quality and scope of the underlying studies that have been gathered (Briggs, 2008; Seidel & Shavelson, 2007; Shavelson & Towne, 2002). In this sense, we harbor no assumptions that this or any other meta-analysis will definitively answer the questions raised about a particular instructional approach. A concern about meta-analysis that we have tacitly raised in this study is that there may often be unexplored issues with regard to the construct validity of the treatment or intervention (in this case inquiry-based teaching). Nonetheless, this same concern could be raised with our attempt to establish a more refined coding approach since, obviously, we were not able to retrospectively attend these classrooms to determine for ourselves the extent to which any of the dimensions of inquiry were or were not enacted and instead had to rely on descriptions of the actual instruction for our process of coding, an inherent limitation in this necessarily retrospective research synthesis.

Furthermore, we experienced a tension between better defining the construct and the size of the sample of studies. On the one hand we wanted a better definition of the construct to put the studies into smaller bins that would reflect more accurately what students experienced in the studies. However, at the same time, by sorting into smaller categories we reduced the number of studies in each bin, making the computations of mean effect sizes across studies more sensitive to any single study with an unusually low or high effect size. This tension brought into sharp relief a tradeoff that must be navigated when conducting any meta-analysis.

Implications for Science Teaching and Learning

The results of this meta-analysis have several implications for the field of reform-oriented science teaching and learning. Overall, the results indicate a positive effect of inquiry-based teaching reforms on student learning of science. Second, results suggest the importance of the role of the teacher in actively guiding student activities in the context of inquiry learning. The current move in science education is away from using inquiry as a catch-all term for reform and toward more detailed descriptions of the practices in which the teacher and student engage (e.g., BOSE, 2012; NRC, 2007). This meta-analysis, by taking the approach of disaggregating different domains of a two-dimensional model of inquiry, has helped to identify particular features of those practices that appear to be more strongly linked to improved student learning of science. We discuss these features in the sections that follow.

Model of Inquiry

We found considerable variability among effect sizes when they were considered as a function of the cognitive and guidance dimensions of inquiry. For the 13 studies where the distinguishing cognitive feature of the inquiry treatment was EC, S, or PECS, the mean effect sizes were a modest .19, .11, and .24, respectively. Yet for the 9 studies where the distinguishing cognitive feature was E or PES, the mean effect sizes were about three times larger (.75 and .72).
With respect to our comparison of studies on the basis of model of inquiry, we found that studies contrasting the epistemic domain and the procedural, epistemic, and social domains had the largest effect size. This finding suggests that engaging students in generating, developing, and justifying explanations as part of other science activities is an important element to helping students learn science. This finding also supports multiple lines of research in the past 20 years that have argued for student engagement in evidence and explanation (NRC, 1996, 2001, 2007), as well as new policy directives in the conceptual framework for the new science standards (BOSE, 2012).

The next largest effect size was found for students that contrasted procedural, epistemic, and social. These studies usually had a traditional treatment group that involved students engaging in content only and therefore represent a classic comparison between a content-focused control and a treatment that integrates the other three domains of inquiry. The size of the effect of this subset of studies on student learning supports the contention of the science education community that inquiry-based teaching has a large, positive effect on student learning.

Surprisingly, a large number of studies in our sample ($n=13$) did not contrast any of the four domains of inquiry we defined between the experimental and control groups. Given the stature of one of these studies within the science education community (White & Frederiksen, 1998), this finding suggests that there may be nuances in enactment of inquiry that were not captured by our coding system. However, many other studies that fell into this category studied representations that support student engagement in inquiry, but not inquiry itself, such as 2-D versus 3-D representations of the same content in support of inquiry learning (Kim, 2006) or diagrams to support students’ conceptual understanding (Ugwu & Soyibo, 2004). The presence of so many studies in our sample indicate that many of the experimental studies performed in the 1996–2006 period did not actually study inquiry-based teaching and learning per se, but rather contrasted different forms of instructional scaffolds that did not substantively change the ways in which students engaged in the domains of inquiry, at least as those domains were conceptualized in this study. Future meta-analyses could explore the idiosyncrasies of these instructional scaffolds to better understand the ways in which these technologies can support inquiry-based teaching.

**Model of Teacher Guidance**

Critics of reform-oriented science teaching have argued that minimally guided instructional approaches are not successful in helping students to learn (Kirschner et al., 2006; Mayer, 2004). However, these critics have equated inquiry-based teaching, which is commonly conceived of as being supported by the teacher (Hmelo-Silver et al., 2007), with discovery learning, which has largely been discredited in mainstream science education circles as a means for getting students to learn specific scientific concepts (Shulman & Keislar, 1966). In this study, we considered discovery learning as occupying a position on the same continuum as teacher-guided inquiry, that is, a variation of the same reform, but with much less guidance provided by the teacher.

The subset of 10 studies that contrasted teacher-led reform conditions with traditional instruction had a mean effect size that was more than twice as large (.65) as the 5 studies that contrasted student-led reform with traditional instruction (.25). One might argue that in the conditions structured or guided by the teacher,
students experienced the domains of inquiry more directly than in the student-led conditions; however, in the absence of detailed implementation data, we cannot draw this conclusion. Nonetheless, the evidence from these studies suggests that teacher-led inquiry lessons have a larger effect on student learning than those that are student led.

This finding illustrates an important nuance in the ongoing debate over the impact of reform-based teaching on student learning. Several papers published in recent years have argued for the failure of reform-based teaching featuring student guidance, notably Kirschner et al. (2006), Klahr and Nigam (2004), and Mayer (2004). However, Hmelo-Silver et al. (2007) argued that Kirschner et al. (2006) inappropriately grouped discovery learning with structured inquiry approaches that actually include guidance from the teacher. The results of this meta-analysis affirm this argument that engaging students in guided inquiry contexts does lead to learning gains when contrasted with comparison groups featuring traditional lessons or unstructured student-led activities.

Reform Science Teaching Terminology

Another finding of this analysis is the variety of terms used by the authors of these studies to describe the treatments and controls being contrasted, which made difficult the search and synthesis process. For example, some of the rigorous studies that featured instructional strategies related to inquiry-based teaching and learning did not label themselves as such (e.g., Hardy et al., 2006) while other studies that did label themselves as inquiry provided few details as to how inquiry was actually operationalized in the instruction (Pine et al., 2006). Furthermore, the different terms used to describe inquiry sometimes refer to the amount of guidance provided by the teacher (e.g., discovery learning) and other times refer to the cognitive activities of the student (e.g., model-based inquiry).

We suggest that greater consistency in keywords would make studies more accessible through search engines, and editors and reviewers could help authors more accurately select keywords for their work. The greater problem of inconsistency in reform science teaching jargon might be addressed through sustained discussion of what all of these terms mean, focusing on teaching and learning practices rather than labels alone. Furthermore, experimental and quasi-experimental studies might benefit by greater consistency and transparency in the type of instruction students experience in the control groups, not only to facilitate future research syntheses, but also to promote consensus within the field as to which approaches to science instruction are most effective.

Conclusion

In closing, we will summarize what we believe are three key findings of this meta-analysis. First, our synthesis of a historical sample of studies from the years in which inquiry-based teaching was returning to prominence indicates a positive effect of this teaching approach on student learning, with a particularly large effect of students engaging in the epistemic domain of inquiry and the procedural, epistemic, and social domains combined. Second, the meta-analysis also indicates higher effect sizes for studies that involved teacher-led activities. Third, and extending beyond the domain of inquiry-based teaching, this meta-analysis has
illustrated how a refined model for an instructional approach can yield more nuanced interpretations of the effects of that approach on student learning.

**Note**

This research was initiated while Erin Marie Furtak was a visiting scholar at the Leibniz-Institute for Science and Mathematics Education (IPN) in Kiel, Germany, with the support of a German Chancellor Fellowship from the Alexander von Humboldt Foundation. The authors thank Manfred Prenzel for his support of this research project while Director of the IPN, Rich Shavelson and Katherine Eggers for reading and providing feedback on multiple drafts of this manuscript, as well as the editors and reviewers who provided helpful guidance in improving its quality and readability.

1. When comparing the three approaches to calculating effect size, we found that Equation 3 yielded the lowest effect size and variance as follows: Glass’s Δ: .70 (SD = .86), ESPooled_SD_Pre = .65 (SD = .73), ESSControl_SD_Post = .50 (SD = .56).

**References**


326


327


**Authors**

ERIN MARIE FURTAK is assistant professor of education with a specialization in science education in the Department of Curriculum and Instruction at the University of Colorado at Boulder, 249 UCB, School of Education, Boulder, CO 80309; e-mail: erin.furtak@colorado.edu. A former high school science teacher, she received her PhD in curriculum and teacher education from Stanford University. Her research explores issues pertaining to the training and retention of secondary science teachers, and the improvement of student learning through reform-based science teaching.

TINA SEIDEL is the Friedl Schöller Chair of Teaching and Learning Research at TUM School of Education, Technische Universität München, Germany. She finished her PhD with excellence at the Leibniz-Institute for Science and Mathematics Education in Kiel, Germany. She is the author of numerous books and articles on the topics of classroom research, teacher and teaching effectiveness, teacher learning, and teacher professional development. Her special interest is in video-based analysis of classrooms, teaching adaptivity and aptitude-treatment-interaction and development of multimedia-based tools in teacher research.

HEIDI IVERSON is a postdoctoral researcher at the School of Education and Human Development at the University of Colorado, Denver. She received her PhD in curriculum and instruction with an emphasis on science education from the University of Colorado at Boulder. Her research interests center around the development and evaluation of innovative approaches to teaching in the college environment, specifically in physics education, and more broadly on innovative approaches to instruction and assessment in science education.

DEREK C. BRIGGS is associate professor and chair of the Research and Evaluation Methodology Program at the University of Colorado at Boulder, where he also serves as an associate professor of quantitative methods and policy analysis. He received his PhD from the University of California at Berkeley. In general, his research agenda focuses upon building sound methodological approaches for the valid measurement and evaluation of growth in student achievement. His daily agenda is to challenge conventional wisdom and methodological chicanery as they manifest themselves in educational research, policy, and practice.