AERA Open
April-June 2018, Vol. 4, No. 2, pp. 1–23
DOI: 10.1177/2332858418767402

Is Inquiry Science Instruction Effective for English Language Learners? A Meta-Analytic Review

Gabriel Estrella
Jacky Au
Susanne M. Jaeggi
Penelope Collins

University of California, Irvine

Despite being among the fastest growing segments of the student population, English Language Learners (ELLs) have yet to attain the same academic success as their English-proficient peers, particularly in science. In an effort to support the pedagogical needs of this group, educators have been urged to adopt inquiry approaches to science instruction. Whereas inquiry instruction has been shown to improve science outcomes for non-ELLs, systematic evidence in support of its effectiveness with ELLs has yet to be established. The current meta-analysis summarizes the effect of inquiry instruction on the science achievement of ELLs in elementary school. Although an analysis of 26 articles confirmed that inquiry instruction produced significantly greater impacts on measures of science achievement for ELLs compared to direct instruction, there was still a differential learning effect suggesting greater efficacy for non-ELLs compared to ELLs. Contextual factors that moderate these effects are identified and discussed.

Keywords: *English Language Learner, science education, inquiry instruction, achievement gap, quantitative research synthesis*

The Next Generation Science Standards (NGSS; Achieve, 2012) reflect the importance of introducing children to scientific and engineering practices early to prepare them for STEM careers. The NGSS framework emphasizes the use of rich content and practices that refine and deepen science inquiry in ways that go beyond the use of hands-on, constructivist approaches to science instruction (Achieve, 2012). However, implementing these standards in elementary school presents unique challenges to educators who must increasingly teach complex concepts and reasoning for English Language Learners (ELLs), or students who have yet to fully develop proficiency in English (Saunders & Marcelletti, 2013). Furthermore, engaging in the science and engineering practices are language intensive for all students and ELL students in particular (Lee, Quinn, & Valdés, 2013). Although the population of ELL students has increased substantially in recent years, their achievement in science has not (Maerten-Rivera, Myers, Lee, & Penfield, 2010; National Center for Education Statistics [NCES], 2014).

To better support the pedagogical needs of this growing population, educators have been encouraged to adopt inquiry-based approaches based on the premise that hands-on instruction makes science learning more engaging, concrete, and meaningful (Janzen, 2008; National Research Council [NRC], 2012; Roseberry & Warren, 2008). Whereas inquiry-based instruction has been shown to improve the science achievement of English-proficient (or non-ELL) students (Furtak, Seidel, Iverson, & Briggs, 2012), consensus regarding both the effectiveness and appropriateness of this approach for ELL students has yet to be established. Substantive differences in the linguistic backgrounds, academic experiences, and pedagogical needs of ELL and non-ELL students have led to disagreement regarding the benefits of inquiry-based approaches for linguistically diverse students. Thus, we seek to conduct a meta-analysis examining the effectiveness of inquiry-based instruction for ELL students. We begin with a brief overview of ELL students' performance in science, the rationale behind teaching ELL students with inquiry-based instruction, and the promises and challenges associated with its application.

The Need for Effective Science Instruction: Underachievement of ELL Students in Science

ELL students' educational attainment has received growing attention due to persistently low achievement in general and in STEM in particular (Bravo & Cervetti, 2014; Diamond, Maerten-Rivera, Rohrer, & Lee, 2014; Lara-Alecio et al., 2012). Despite increased resources to enhance STEM education, ELL students have yet to attain the same

level of academic success as their English-proficient peers (Lee & Buxton, 2013; Maerten-Rivera et al., 2010; Tong, Irby, Lara-Alecio, & Koch, 2014). For example, ELL students consistently score lower on the science portion of the National Assessment of Educational Progress (NAEP) at all grade levels and are more likely to score below basic (NCES, 2014). These findings indicate that ELL students are in need of greater support in STEM education as compared to their non-ELL peers (Genesee, Lindholm-Leary, Saunders, & Christian, 2005; Goldenberg, 2013).

ELL students' low achievement in science may be attributed in part to their limited proficiency in English and weak mastery of academic language (Kieffer, Lesaux, Rivera, & Francis, 2009). Scientific texts are linguistically complex, informationally dense, and highly technical (Echevarria, Richards-Tutor, Canges, & Francis, 2011; Fang, 2006). The linguistic complexity of scientific texts can impede meaningful learning for ELL students by interrupting information processing and conceptual understanding (Fang, 2006; Janzen, 2008). Thus, ELL students' science learning may be constrained by their proficiency in English (Lee, 2005).

Potential Benefits of Learning Science With Inquiry Instruction

One view is that ELL students learn best when instruction is situated within meaningful, interactive activities that leverage the language and cultural backgrounds of students (Bravo & Cervetti, 2014; Echevarria et al., 2011). Inquiry instruction is grounded in the constructivist principle that meaningful learning occurs when students engage in authentic activities that promote active knowledge construction through self-guided exploration (Bruner, 1996; Lee, 2005). Students are encouraged to construct knowledge by posing questions about the natural world, test theories through carefully planned investigations, and draw conclusions based on empirical results (Bruner, 1996). Thus, teachers facilitate meaningful dialogue, experimentation, and engagement (Minner, Levy, & Century, 2010). Inquiry instruction is often contrasted with traditional approaches, such as direct instruction, which aim to build factual knowledge through explicit exposition and highly structured teacher guidance (Kirschner, Sweller, & Clark, 2006). Although direct instruction is commonly used, inquiry learning has been found to improve students' attitudes toward science (Jiang & McComas, 2015), enhance problem-solving skills (Lazonder & Harmsen, 2016), and increase learning outcomes (Alfieri, Brooks, Aldrich, & Tenenbaum, 2011).

Although most examinations of inquiry instruction to date involve non-ELL students, its benefits are assumed to generalize to ELL students in a number of ways. First, inquiry instruction's use of engaging, multisensory activities is assumed to increase ELL students' access to scientific content by reducing the demands of scientific language (Janzen, 2008). Second, its multimodal nature encourages physical and cognitive engagement to support deeper levels of learning (Huerta & Jackson, 2010). Third, inquiry instruction encourages ELL students to communicate their understanding of scientific concepts and procedures, which may promote their oral and written language skills (August, Branum-Martin, Cardenas-Hagan, & Francis, 2009). Finally, the collaborative nature of inquiry instruction is thought to promote rich learning experiences for ELL students that foster both conceptual knowledge and scientific communication (Lee & Buxton, 2013).

Concerns Regarding the Effectiveness of Inquiry Instruction for ELL Students

Despite its potential benefits, there remain concerns and contradictory findings regarding the effectiveness of inquiry instruction with ELL students. First, ELL students may lack sufficient English proficiency to benefit fully from inquiry instruction (August et al., 2009; Bresser & Fargason, 2013; Huerta, Tong, Irby, & Lara-Alecio, 2016). Despite using multimodal approaches to pedagogy, inquiry instruction still has heavy linguistic demands, requiring students to generate predictions, communicate their findings, and engage in meaningful scientific discourse. However, many ELL students are still developing the very language skills critical for active participation and building understanding of the content. Thus, the provision of more hands-on, active learning opportunities may not sufficiently address the linguistic challenges faced by ELL students in the science classroom (August et al., 2009; Bravo & Cervetti, 2014; Lee, Deaktor, Enders, & Lambert, 2008).

Second, the assumption that inquiry instruction is more effective than traditional methods has also been challenged (e.g., Kirschner et al., 2006; Tobias & Duffy, 2009). The hands-on, self-guided exploration characteristic of inquiry may not provide sufficient instructional guidance and structure to facilitate meaningful learning and transfer (Mayer, 2004). Although hands-on instruction may provide students with salient, highly contextualized learning experiences, inquiry instruction may not provide enough of a framework to enable students to represent scientific principles and understanding more abstractly and generalize what they have learned to new contexts.

Finally, the benefits of inquiry instruction may be limited to students who already have sufficient prior knowledge to support exploratory learning (Kirschner et al., 2006; Klahr & Nigam, 2004). Because ELL students' access to quality instruction is often limited by English-only instruction, tracking into remedial classes, and attending English support services at the exclusion of content-area instruction (Robinson-Cimpian, Thompson, & Umansky, 2016), they may lack the academic preparation to fully benefit from inquiry instruction. Thus, the effects of inquiry instruction for ELL students requires greater examination.

Factors That May Influence the Effectiveness of Inquiry Instruction

From a developmental perspective, there are compelling reasons to expect that the effectiveness of inquiry-based instruction may differ on the basis of student grade level (Meyer, 2000). One factor is that as ELL students progress from first grade and beyond, they build their knowledge base in science, proficiency in English, and metacognitive abilities—all of which contribute to higher learning and achievement. Consequently, inquiry-based instruction may be more advantageous for older ELL students who, compared to their younger counterparts, are more likely to have the requisite skills and knowledge to meet the demands of learning science with inquiry-based instruction. On the other hand, the increasingly rigorous academic and linguistic demands associated with science inquiry in higher grade levels might overburden older ELL students and result in diminished learning (Tolbert Stoddart, Lyon, & Solis, 2014).

Second, the effectiveness of inquiry instruction may be influenced by factors such as teacher preparation and instructional time. Many elementary school teachers report they have been inadequately prepared to teach ELL students science (Cervetti, Kulikowich, & Bravo, 2015; Zwiep & Straits, 2013). However, teachers' instructional skills and pedagogical knowledge have been shown to have a significant impact on students' science achievement (Heller, Daehler, Wong, Shinohara, & Miratrix, 2012). Professional development has been found to improve the delivery of inquiry instruction by raising teachers' pedagogical knowledge and understanding of ELL students' learning needs (Yoon, Duncan, Lee, Scarloss, & Shapley, 2007).

Third, inquiry instruction requires heavy investments in instructional time. There is considerable variation in the amount of class time devoted to inquiry instruction, which may also influence its effectiveness for ELL students (Baker, Fabrega, Galindo, & Mishook, 2004; Dorph, Shields, Tiffany-Morales, Hartry, & McCaffrey, 2011). Thus, our meta-analysis considers professional development and instructional time in a moderation analysis.

Prior Reviews of Inquiry-Based Instruction for ELL and Non-ELL Students

Several narrative reviews summarizing the prevailing state of knowledge on effective teaching approaches with ELL students provide initial support for the use and effectiveness of inquiry-based instruction with ELL students. Lee (2005) performed a systematic review of research on the science education (K–12) of ELL students and found that hands-on, inquiry-based instruction was generally associated with positive achievement outcomes among all students, including those with lower levels of English proficiency and prior science experience. More recently, Janzen's (2008) narrative review on content-area instruction in science with ELL students found similar evidence suggesting that inquiry-based instruction led to improvements in both ELL students' language development and science achievement. Although these reviews offer a useful summary of research on the effectiveness of inquiry-based instruction with ELL students, they use qualitative rather than quantitative methods, do not provide effect size estimates and furthermore, are not the most current anymore.

Three more recent meta-analyses comparing the effectiveness of inquiry-based instruction with direct instruction support the advantage of inquiry-based instruction. First, Alfieri et al.'s (2011) meta-analysis contrasted the effectiveness of direct instruction to both guided and unguided forms of inquiry-based instruction. They found that inquiry-based instruction produced greater achievement outcomes in science than direct instruction (d = .11). Similarly, Furtaket al.'s (2012) meta-analysis found that inquiry-based instruction resulted in significantly greater learning outcomes (d = .50). Finally, Lazonder and Harmsen (2016) showed that guided forms of inquiry instruction produced a positive effect on students' science content knowledge (d = .50) and ability to perform inquiry (d = .66). Although these meta-analyses provide evidence suggesting that inquiry-based instruction can be an effective method of learning for students as compared with traditional instruction, they were based on studies conducted primarily with mainstream English-proficient students, and thus, their results may not generalize to ELL students.

Present Study

Previous syntheses of research have concluded that inquiry-based instruction is a particularly effective approach for improving the science achievement outcome for students. However, to our knowledge, no study to date has explicitly evaluated changes in ELL students' science achievement as a result of receiving inquiry instruction in a comprehensive and quantitative synthesis. To this end, we conducted a meta-analysis to determine the extent to which inquiry instruction serves ELL students' science achievement, addressing the following questions:

- *Research Question 1*: Is inquiry-based science instruction an effective method of teaching for ELL students relative to direct instruction?
- *Research Question 2*: Does inquiry science instruction provide comparable learning benefits to ELL students relative to their English proficient peers?
- *Research Question 3*: What types of factors, if any, moderate the impact of inquiry instruction on science achievement outcomes for ELL students?

Method

Selection of Studies and Data Collection

Inclusion criteria. We developed selection criteria that would capture empirical studies designed to evaluate the impact of inquiry instruction on science achievement for ELL students. Both published and unpublished studies were eligible to be included as long as they (a) used an experimental or quasi-experimental research design, (b) provided data for ELL students between kindergarten and sixth grade, (c) included a treatment that received inquiry instruction and either a business-as-usual control receiving direct instruction or a non-ELL student comparison group, (d) assessed the effects of inquiry instruction on students' science learning outcomes and reported these effects quantitatively, (e) provided sufficient data to calculate effect sizes, and (f) were either published or translated in English. To avoid sample bias to the best extent possible, studies that focused exclusively on students who were reclassified as fluent English proficient (i.e., former ELLs) were excluded from this meta-analysis. Furthermore, studies that combined results for ELL and non-ELL students or elementary and non-elementary school students such that effect sizes could not be extracted independently for each subsample were also excluded.

Search procedure. A comprehensive and systematic search was conducted (between 2000 and 2016) using ERIC, PsycINFO, and Google Scholar, with the search terms *science, instruction, education, teaching, K–6, methods, English as a second language, English language learner, limited English proficient, inquiry, discovery, hands-on*, and *projects strategies*. The search was restricted to studies that were published in the years 2000 to 2016. To identify unpublished studies in ERIC and Google Scholar, we modified the search parameters to include dissertations, theses, and conference proceedings. We also submitted our selected articles to both forward and backward searches. Forward searches were carried out by searching for articles that cited other studies that met our search criteria, while backward searches were conducted by manually reviewing the reference sections of each paper for additional studies that matched our search criteria. Studies identified in literature reviews and prior syntheses were also reviewed for inclusion. Finally, we contacted authors of the included studies to solicit other published or unpublished studies that may be relevant to this meta-analysis.

Study selection. This search procedure returned over 5,000 potentially relevant articles. Using the selection criteria established previously, we examined the title, abstract, and keywords of each article. Studies that met the most fundamental aspects of the selection criteria—that is, whether or not a study investigated the effect of (a) inquiry-based instruction on the (b) science achievement of (c) ELL students—were flagged for potential inclusion and saved for a second review. When abstracts did not provide adequate information for eligibility judgments, the full text of the article was obtained and screened for potential inclusion using the aforementioned search criteria. If multiple reports of the same study were identified (e.g., dissertation/thesis, journal article), they were grouped together and cross-referenced for complete information, and the most comprehensive study was retained. Based on this first round of the literature search, 32 articles were flagged as potentially relevant.

In the second round of reviews, we evaluated each article in greater detail. Six studies were excluded because they lacked an eligible treatment/comparison group or science achievement measure or did not provide sufficient information to calculate effect sizes. Studies with missing effect size information were excluded only if we could not obtain the data to estimate effect sizes after requesting them from the corresponding authors. Disagreements regarding whether to include a study were discussed by the research team until consensus was reached. Overall, this selection procedure yielded a total of 26 studies for inclusion in the meta-analysis. Figure 1 summarizes the study search procedure and selection criteria.

Coding of studies. First, students were classified based on their English proficiency (Saunders & Marcelletti, 2013). Students who were non-native speakers of English with limited English proficiency were coded as *ELL*, and native English speakers and language-minority students proficient in English were coded as *non-ELL*. Instruction involving hands-on, self-guided learning tasks requiring students to construct science knowledge using questions and investigations was coded as *inquiry instruction* (Bruner, 1996; Furtak et al., 2012). Explicit instruction using highly structured lectures, demonstrations, textbooks, or other teacher-centered methods was coded as *direct instruction* (Alfieri et al., 2011; Mayer, 2004). Finally, *science achievement outcomes* were coded if they quantitatively assessed changes in students' performance on measures of conceptual, factual, or procedural knowledge (Minner et al., 2010).

Implementation variables were coded for moderation analysis. We coded the *length of the intervention* as the number of months of intervention. *Student grade level* was coded for K through six. *Professional development training* was coded as 1 when it was provided and 0 when no training was provided. When professional development was provided, the *duration* in hours was coded, and the *dosage* was categorized as small if under 15 hours were

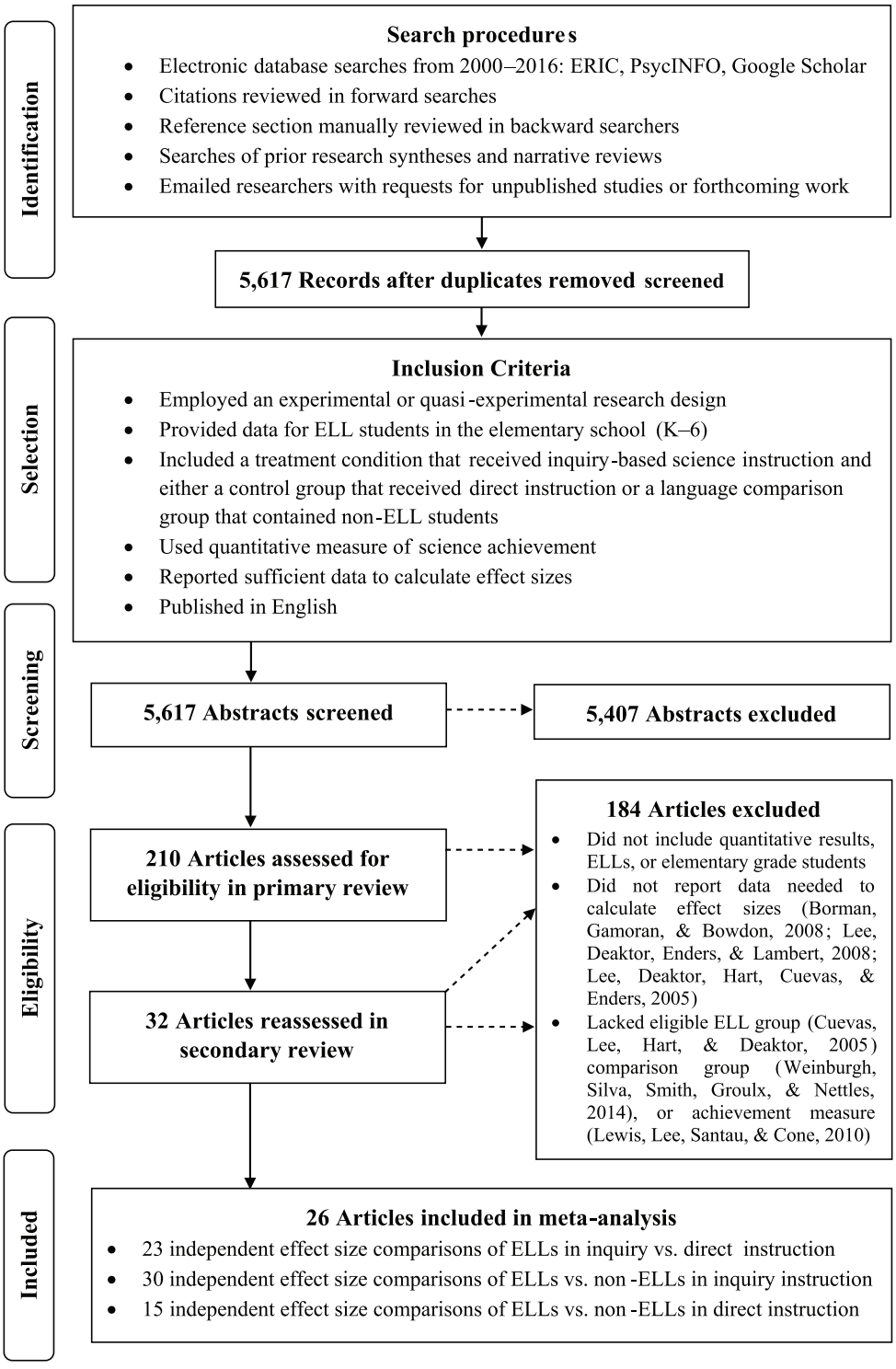

FIGURE 1. *Flow diagram of study selection procedure and selection criteria.*
Note. ELL = English language learner.

provided or large if 15 or more hours were provided. The *focus of the training regime* was considered ELL focused when training addressed the needs or instruction of ELL students. It was coded as non-ELL focused if training was not specific to the needs of ELL students, such as addressing science pedagogy in general.

The methodological features of each study were coded based on its *publication status* (published journal article vs. unpublished dissertation/technical report), *research design* (randomized experiment vs. quasi-experiment), *measurement design* (pretest and posttest vs. posttest only), *assessment format* (multiple choice vs. constructed response), and *assessment type* (researcher-developed test vs. standardized test).

The process of coding was conducted by the first author using a standardized coding protocol developed in advance by the research team (available on request from the authors). However, to ensure reliability and accuracy, all studies were double-coded by the second author. Interrater reliability was established by calculating the percentage of overlap between each coder, which yielded a high percent agreement of 93.6%. Coding discrepancies were discussed as a group until consensus was reached.

Meta-Analytic Procedures

Evaluating the effects of inquiry instruction for ELL students. We derived three separate meta-analytic effect size (ES) estimates based on standardized mean differences. First, we evaluated the effectiveness of inquiry instruction, or treatment ES, using the standardized mean difference in science achievement outcomes between ELL students who learned with inquiry instruction (treatment condition) and ELL students who learned with traditional instruction (control condition). Positive values for treatment ES indicate that ELL students in the treatment condition outperformed those in the control condition.

Examining the effects of inquiry instruction between ELL and non-ELL students. The second analysis examined whether inquiry instruction had similar benefits for ELL and non-ELL students. We calculated the inquiry ES using the standardized mean difference in learning outcomes between ELL and non-ELL students who received inquiry instruction within studies that reported data for both groups. Positive values for inquiry ES indicate that ELL students showed greater gains in inquiry instruction than non-ELL students.

To contextualize the inquiry ES findings, we estimated the effect size for traditional science instruction (traditional ES) using the standardized mean difference in learning outcomes between ELL and non-ELL students who received traditional instruction within studies that reported data for both groups. Positive values for traditional ES indicate that ELL students showed greater gains with traditional instruction than non-ELL students. Studies that reported information to estimate one effect size but not another were included in all analyses for which sufficient information was provided.

Computation of effect sizes. The calculation of the standardized mean difference (Cohen's *d*) effect size was estimated depending on the data provided. First, when only posttest data were available, the standardized mean difference was calculated:

$$d = \frac{\bar{X}_T - \bar{X}_C}{SD_{pooled}},$$

where $\bar{X}_T$ is the mean posttest score of the treatment condition, $\bar{X}_C$ is the mean posttest score of the control condition, and SD_{pooled} is the pooled standard deviation (Lipsey & Wilson, 2001). The pooled standard deviation was calculated as follows:

$$SD_{pooled} = \sqrt{\frac{(n_T - 1)SD_T^2 + (n_c - 1)SD_C^2}{n_T + n_C - 2}},$$

where n_T and n_C are the sample sizes associated with the treatment and control group, and SD_T and SD_C reflect their respective standard deviations.

When pretest and posttest data were reported for both groups, pretest-adjusted estimates of the standardized mean difference were calculated as:

$$d = \frac{(\bar{X}_{Tpost} - \bar{X}_{Tpre}) - (\bar{X}_{Cpost} - \bar{X}_{Cpre})}{SD_{pooled}},$$

where $\bar{X}$ is the mean test score for the treatment (*T*) and control condition (*C*) measured before (*pre*) and after (*post*) the learning phase, and SD_{pooled} is the pooled standard deviation of pretests (Morris, 2008). This method of estimation was preferred as it produces more conservative effect size estimates by adjusting for baseline differences in prior knowledge (Furtak et al., 2012).

When dichotomous data were reported (e.g., proportion of students who attained proficiency on a standardized test), we converted the log odds ratio of successes between groups into the standardized mean difference using the transformation procedure outlined by Borenstein, Hedges, Higgins, and Rothstein (2009).

If means and standard deviations were missing but regression coefficients reported, effects sizes were approximated using the *t* statistic corresponding to the null hypothesis of independent group differences between the treatment and control condition (Borenstein et al., 2009; Lipsey & Wilson, 2001):

$$d = t\sqrt{\frac{n_T + n_C}{n_T n_C}},$$

where n_T and n_C are the sample sizes for the treatment and control conditions and *t* test value for the group comparison. However, because estimates derived from multivariate regression analyses yield partial effect sizes that may not be comparable across studies (Becker & Wu, 2007), sensitivity

analyses were performed to examine whether the results where robust to their inclusion.

Finally, to adjust for upward bias in Cohen's *d* associated with small samples ($n < 20$), all effect sizes were transformed into Hedge's *g* using the small-sample correction factor proposed by Hedges (1982):

$$g = \left[1 - \left(\frac{3}{4(n_T + n_C) - 9}\right)\right] * d,$$

where n_T and n_C are the respective sample sizes for the treatment and control conditions, and *d* is the original standardized mean difference effect size. All effect size computations and subsequent analyses were conducted using the software Comprehensive Meta-Analysis (CMA), version 3 (Borenstein et al., 2014), unless otherwise noted.

Dependent effect sizes. To resolve statistical dependence among studies reporting multiple outcomes for the same group of students, we report the mean effect size for all outcomes to yield a single effect size per study. Similarly, for longitudinal studies involving the same cohort of students, effect sizes were collapsed together to yield a single average effect size per study. We report the mean effect size for multiple treatment groups when they were compared to a single control group. However, effect sizes generated from two or more different subgroups (i.e., grade levels, cohorts of students, or treatments) within a study such that each subgroup was accompanied with its own distinct comparison group were treated as independent (Borenstein et al., 2009). This made it possible for multiple effect sizes to be extracted from a single study. These procedures ensured that each effect size was estimated based on an independent set of data and that each analysis was conducted with an independent set of effect sizes.

Data synthesis. To estimate the overall effect size, studies were issued weights based on their level of precision (i.e., standard error). Because the effects of inquiry instruction on science achievement outcomes were assumed to vary among studies as a function of population, intervention, and methodological differences, we used random effects models to calculate the overall weighted mean effect size ($\bar{g}$) as:

$$\bar{g} = \frac{\sum(w_i * g_i)}{\sum w_i},$$

where g_i is the observed effect size for study$_i$ and w_i is the inverse variance weight assigned to study$_i$ (Borenstein et al., 2009). This approach allows relatively greater weight to be assigned to studies with higher levels of precision.

Heterogeneity of effect sizes. We used the *Q* test of heterogeneity to examine the variation in effect size estimates between studies (Lipsey & Wilson, 2001). Moreover, the I^2 statistic quantifies the percent of variation attributable to true heterogeneity relative to sampling error (Higgins, Thompson, Deeks, & Altman, 2003). Overall, I^2 values range from 0% to 100%, with increasing values reflecting greater levels of heterogeneity.

Moderation analysis. When there was significant heterogeneity across studies, we conducted moderation analyses to examine whether variation among effect sizes could be explained by factors that differ between studies. For categorical variables, we performed a *Q* test of between-group differences (Q_B) using CMA's one-way ANOVA function. For continuous variables, we tested the relation between a moderator and magnitude of effect size using CMA's unrestricted maximum likelihood meta-regression function. All moderation analyses were conducted using random effects models weighted by the inverse variance of effect sizes.

Sensitivity Analyses and Robustness Checks

Four sensitivity analyses were used to assess the impact of statistical methods and data inclusion choices on the conclusions of the results and therefore examine the robustness of our findings.

Robust variance estimation. As noted previously, we resolved statistical dependence among our sample of effect sizes using standard meta-analytic methods, namely, collapsing multiple effect sizes across studies to create a single synthetic effect. To utilize all effect sizes from each study, we reanalyzed the data set of effect sizes using robust variance estimation (RVE; Hedges, Tipton, & Johnson, 2010) with a correction for small sample size bias (Tipton, 2015). This approach permits the synthesis of multiple dependent effect sizes by adjusting the standard errors to account for an assumed correlation (ρ) between effect sizes within studies, thereby minimizing the loss of information that occurs through aggregation. One important limitation to this approach, however, is that a minimum of 40 independent studies with an average of five effect sizes per study are needed to estimate a meta-regression coefficient (Tanner-Smith & Tipton, 2014). This issue is particularly problematic in meta-analyses involving categorical variables with multiple levels. As a result, RVE methods were employed in the synthesis of overall weighted mean effect sizes. All analyses using RVE were conducted in the R statistical environment (version 3.4.2) using the *robumeta* package (Fisher & Tipton, 2014; Tanner-Smith & Tipton, 2014).

Outliers. Boxplots were used to identify potential outliers, defined as effect sizes that were 1.5 interquartile ranges above the 75th percentile range or below the 25th percentile

range of the distribution. Two effect sizes were identified as outliers in the treatment ES analyses. Because these outliers could not be attributed to methodological or theoretical differences between each study, coupled with the relatively small number of studies in the sample, we elected not to eliminate these estimates. Rather, we adjusted the outliers downward to more conservative values using the 90% Winsorization procedure described by Lipsey and Wilson (2001). All analyses were subsequently carried out using the adjusted data set. Results for the original sample are reported in a sensitivity analysis.

Study quality. We assessed the methodological quality of included studies using a version of the Quality Assessment Tool for Quantitative Studies (National Collaborating Centre for Methods and Tools, 2008), which was adapted for use with educational research. This quality appraisal tool uses judgments about the extent to which bias may be present in six methodological domains to produce an overall quality rating of weak, moderate, or strong. Although methodologically rigorous studies are more likely to produce valid results (Higgins, Altman, & Sterne, 2017), we decided not to exclude studies on the bases of methodological quality. Including these studies allowed us to maintain our sample of effect sizes and provides a more complete picture of the current research landscape. However, to examine whether our findings were sensitive to differences in study quality, we conducted a sensitivity analysis that excluded studies with an overall quality rating of weak.

Publication bias. We assessed the potential for publication bias among the sample of effect sizes included in the treatment ES analysis as studies that report null findings or relatively small effects are less likely to be published (Rosenthal, 1979; Song, Hooper, & Loke, 2013). We tried to mitigate publication bias a priori by seeking to include unpublished work ($k = 6$). The extent and impact of publication bias was assessed graphically using funnel plots and statistically using Egger's linear regression test (Egger, Smith, Schneider, & Minder, 1997) and a trim and fill analysis of the corresponding funnel plots (Duval & Tweedie, 2000).

Results

Contrasting the Effects of Inquiry and Traditional Instruction for ELL Students

Our first objective was to evaluate whether inquiry-based instruction is more effective than traditional instruction for ELL students. To address this question, we tested if inquiry instruction is more effective than traditional instruction for ELL students by calculating the standardized mean difference ($k = 23$) in science learning outcomes between ELL students who received inquiry instruction ($n = 4{,}204$) and ELL students who received traditional instruction ($n = 4{,}087$). Figure 2 shows that overall, ELL students receiving inquiry instruction tended to obtain science scores that were over one-quarter a standard deviation higher than those receiving traditional instruction, Treatment ES = + 0.28 ($SE = 0.07$, $p < .001$). The 95% confidence interval ranges from 0.15 to 0.41, suggesting that overall inquiry instruction produces a small positive impact on ELL students' learning outcomes.

Contrasting the Effects of Inquiry and Traditional Instruction Between Language Groups

Our second question examines whether inquiry instruction leads to comparable learning benefits to ELL and non-ELL students and are presented in Figure 3. To this end, we estimated the standardized mean difference ($k = 30$) in science learning outcomes between ELL ($n = 5{,}459$) and non-ELL ($n = 42{,}700$) students receiving inquiry instruction. The significant inquiry ES of −0.31 ($SE = 0.08$, $p < .001$) suggests that non-ELL students obtained science achievement scores that were about one-third a standard deviation higher than those of ELL students.

Next, we investigated how the achievement gap between ELL and non-ELL students receiving inquiry science instruction compared to relative performance of ELL and non-ELL students receiving traditional science instruction. To do so, we calculated the standardized mean difference ($k = 15$) in science learning outcomes between ELL ($n = 3{,}085$) and non-ELL ($n = 9{,}364$) students who received traditional instruction in the control condition (see Figure 4). Overall, non-ELL students obtained science scores that were almost half a standard deviation higher than those of ELL students in traditional classrooms, traditional ES = −0.46 ($SE = 0.12$, $p < .001$). The achievement gap between ELL and non-ELL students was greater in science classrooms using traditional instruction ($\bar{g} = -0.46$) than in those using inquiry instruction ($\bar{g} = -0.31$). However, a caveat is that the 95% confidence intervals for these effect sizes overlap. Thus, these findings suggest that inquiry instruction may help attenuate the science achievement gap for ELL students.

Heterogeneity of Effect Sizes

Heterogeneity analyses were conducted to the presence and degree between-study variation using the Q-test and I^2 statistic. First, we tested the treatment ES, or the degree to which ELL students obtained higher outcomes with inquiry instruction, for heterogeneity. We found a high degree of heterogeneity among the studies, $Q = 126.84$, $df = 22$, $p < .001$, with the I^2 statistic revealing that 83% of the total observed variance could be attributed to between-study differences rather than within-study sampling error. Next, we examined the heterogeneity of the inquiry ES or the achievement gap between ELL and non-ELL students receiving inquiry instruction. Once again, there was a

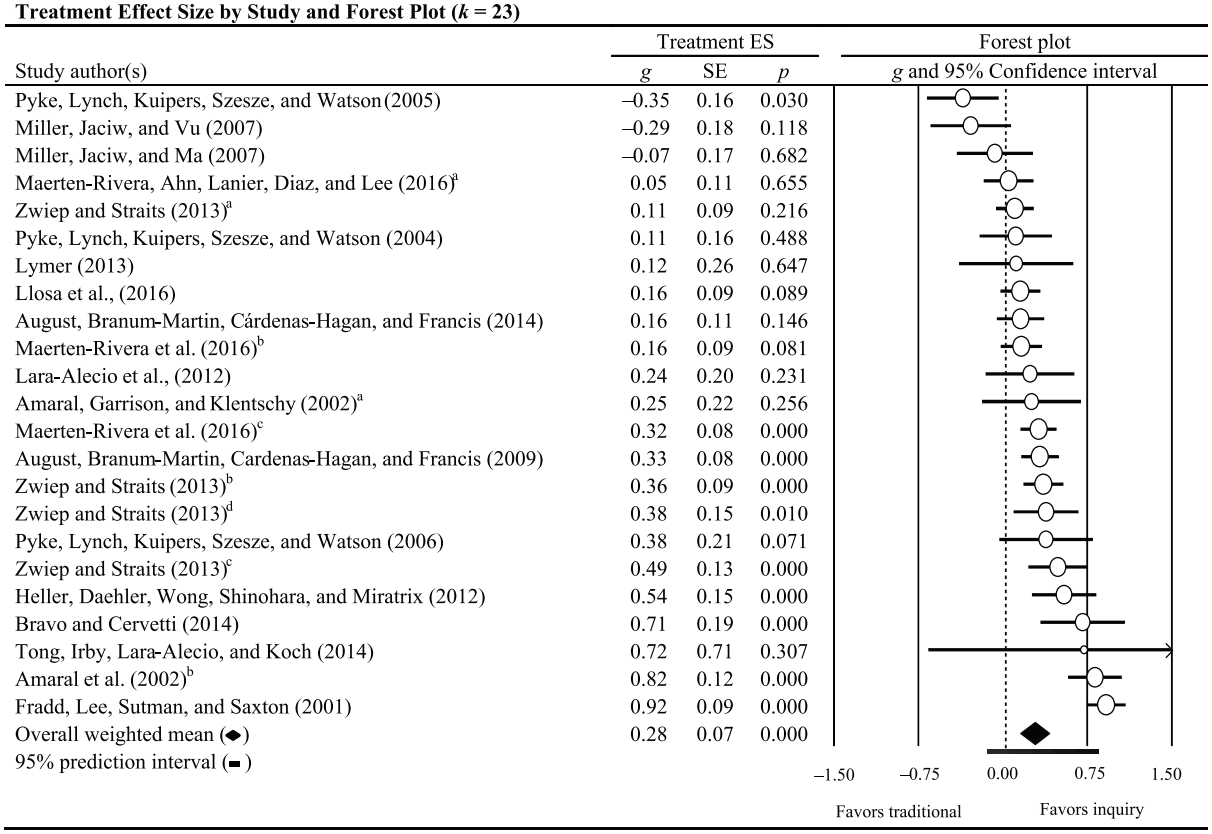

FIGURE 2. *Estimated mean treatment effect size (difference in science achievement between English language learners in treatment and control conditions) for each study with overall mean weighted effect size. Forest plot showing treatment effect sizes with 95% confidence interval and 95% prediction interval. Studies with alphabetic superscripts refer to multiple independent effect sizes generated from the same study.*

high degree of heterogeneity among the studies, $Q = 377.03$, $df = 29$, $p < .001$, with the I^2 statistic indicating that 92% of the variance could be attributed to between-study differences. Finally, the traditional ES, or the achievement gap between ELL and non-ELL students receiving traditional science instruction, also showed a high degree of heterogeneity, $Q = 246.77$, $df = 11$, $p < .001$, with the I^2 statistic revealing that 92% of the variance is attributable to true heterogeneity. Due to the significant heterogeneity across each sample of effect sizes, we conducted a set of moderator analyses across each sample of effect sizes to identify the sources of between-study variation.

Moderation Analyses

To identify moderating factors that may influence the effect of inquiry instruction on ELL students' science achievement, we calculated two sets of analyses to examine the potential influence of categorical and continuous moderators for each effect size. Table 1 presents the results for categorical moderators obtained from the subgroup analyses for treatment ES, while the subgroup moderation results corresponding to traditional ES and inquiry ES are displayed in Table 2 and Table 3, respectively. Table 4 presents the results for continuous moderators obtained from the weighted random effects meta-regression analyses. To mitigate against the potential of confounding variable bias in the meta-regression analyses, each predictor is included in the regression analyses as a covariate, along with the following indicators of methodological quality: publication status, research design, and measurement design.

Publication status. The effect of publication status moderated the findings for treatment ES ($Q_B = 8.19$, $df = 1$, $p = .004$). Studies published in peer-reviewed journals had average treatment ESs that were significantly larger than those in nonpublished studies ($\bar{g} = 0.37$ vs. $\bar{g} = -0.04$). Although a similar pattern of results was observed for both the inquiry ES and treatment ES, such that published studies yielded larger effect sizes than unpublished studies, the between-levels difference was not significant for either of these effect sizes ($p > .10$). These initial analyses suggest that there may be a publication bias for the treatment ES.

Inquiry Effect Size by Study and Forest Plot (k = 30)

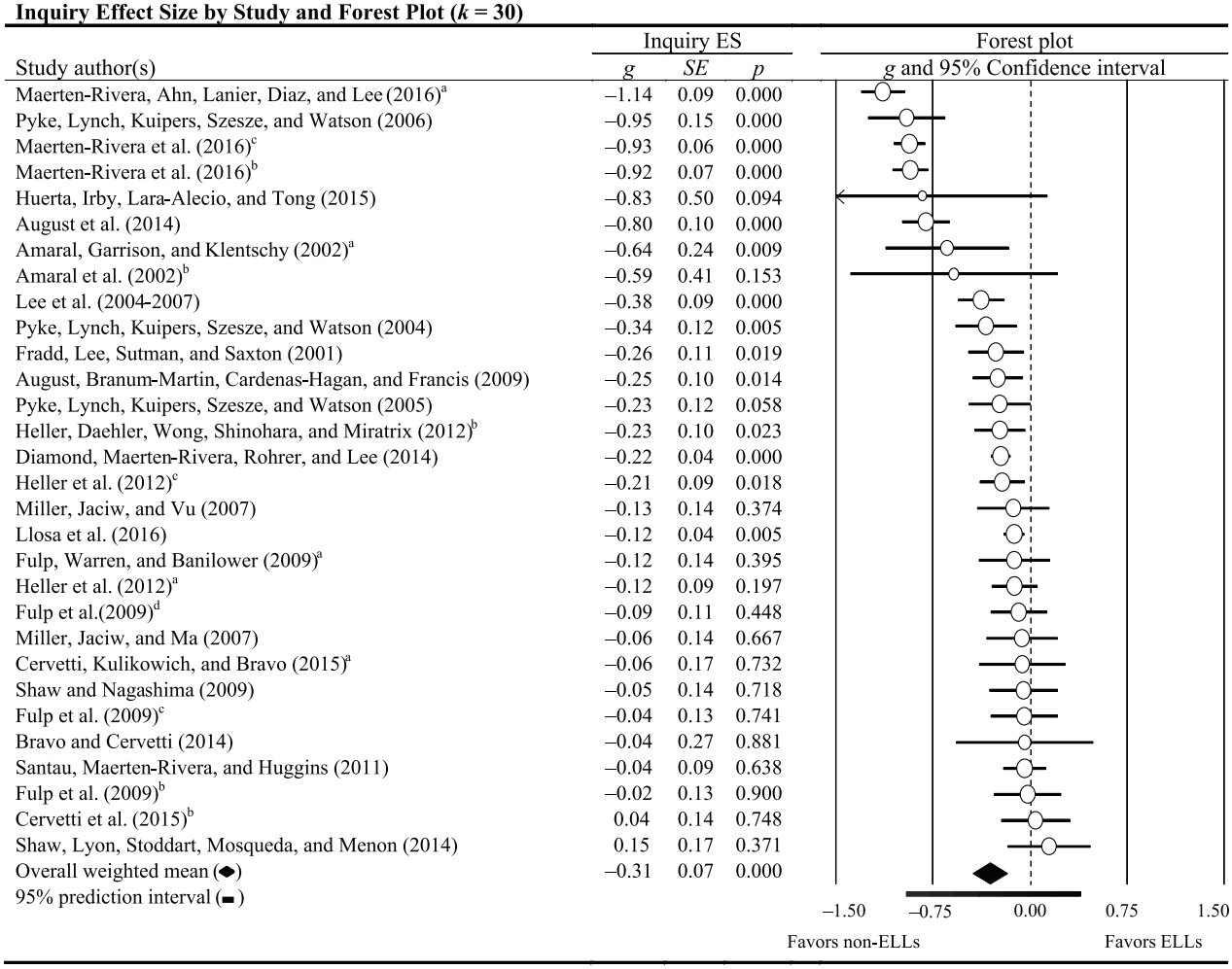

FIGURE 3. *Estimated mean inquiry effect size (difference in science achievement between English language learners and non-English language learners in treatment condition) for each study with overall weighted effect size. Forest plot showing inquiry effect sizes with 95% confidence interval and 95% prediction interval. Studies with alphabetic superscripts refer to multiple independent effect sizes generated from the same study. The data used to calculate an overall effect size for Lee et al. 2004–2007 is based on information reported in Lee, Maerten-Rivera, Penfield, Leroy, and Secada (2008); Lee, Mahotiere, Salinas, Penfield, and Maerten-Rivera (2009); and Lee, Penfield, and Maerten-Rivera (2009).*

Research design. We found significant moderation effects based on the research design for the treatment ES (Q_B = 5.20, df = 1, p = .02) but not for the inquiry ES or traditional ES. Studies using quasi-experimental designs showed significantly larger treatment ESs ($\bar{g}$ = 0.47) than those using randomized experimental designs ($\bar{g}$ = 0.17).

Measurement design. The type of measurement design moderated findings for both the inquiry ES (Q_B = 14.52, df = 1, p < .001) and traditional ES (Q_B = 16.18, df = 1, p < .001) but not for the treatment ES. Studies that used posttest-only designs revealed science achievement gaps between ELL and non-ELL students that were on average three to four times larger than those using pretest-posttest designs for both the inquiry ES ($\bar{g}$ = −0.17 vs. $\bar{g}$ = −0.66) and traditional ES ($\bar{g}$ = −0.17 vs. $\bar{g}$ = −0.66). These findings suggest that ELL and non-ELL students show varying levels of prior knowledge and subsequent growth in science.

Assessment format and assessment type. Whereas assessment format did not moderate the findings for any of the three main effect sizes, differences in assessment type was a moderator for the treatment ES (Q_B = 5.08, df = 1, p = .024), inquiry ES (Q_B = 5.03, df = 1, p = .025), and traditional ES (Q_B = 2.92, df = 1, p = .087). For the treatment ES, studies using researcher-developed assessments ($\bar{g}$ = 0.39) revealed larger gains in science

Traditional Effect Size by Study and Forest Plot (k = 15)

Study author(s)	Traditional ES g	SE	p
Maerten-Rivera, Ahn, Lanier, Diaz, and Lee (2016)[c]	–1.14	0.07	0.000
Maerten-Rivera et al. (2016)[a]	–1.12	0.08	0.000
Pyke, Lynch, Kuipers, Szesze, and Watson (2006)	–1.06	0.16	0.000
Maerten-Rivera et al. (2016)[b]	–1.00	0.07	0.000
Amaral, Garrison, and Klentschy (2002)[b]	–0.71	0.21	0.001
August, Branum-Martin, Cárdenas-Hagan, and Francis (2014)	–0.66	0.09	0.000
Llosa et al. (2016)	–0.34	0.07	0.000
Heller, Daehler, Wong, Shinohara, and Miratrix (2012)	–0.29	0.08	0.000
Amaral et al. (2002)[a]	–0.27	0.20	0.164
August, Branum-Martin, Cardenas-Hagan, and Francis (2009)	–0.18	0.10	0.064
Pyke, Lynch, Kuipers, Szesze, and Watson (2004)	–0.17	0.12	0.158
Miller, Jaciw, and Vu (2007)	–0.11	0.15	0.460
Pyke, Lynch, Kuipers, Szesze, and Watson (2005)	–0.05	0.12	0.698
Bravo and Cervetti (2014)	0.06	0.21	0.774
Miller, Jaciw, and Ma (2007)	0.20	0.15	0.162
Overall weighted mean (◆)	–0.46	0.14	0.001
95% prediction interval (–)			

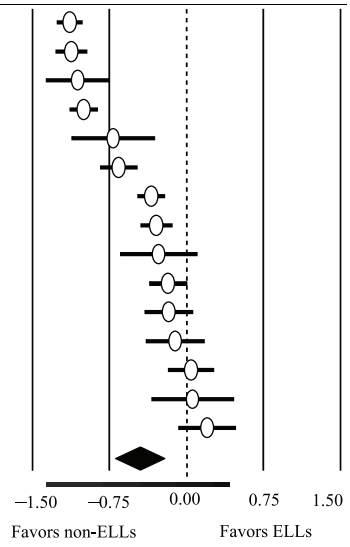

FIGURE 4. *Estimated mean traditional effect sizes (difference in science achievement between English language learners and non-English language learner students in control condition) for each study with overall weighted effect size. Forest plot showing traditional effect sizes with 95% confidence interval and 95% prediction interval.*

achievement than those using standardized assessments ($\bar{g}$ = 0.12). In contrast, studies using standardized assessments revealed greater science achievement gaps between ELL and non-ELL students than those using more proximal, researcher-developed assessments for both the inquiry ES and traditional ES ($\bar{g}$ = −0.56 vs. $\bar{g}$ = −0.24; $\bar{g}$ = −0.69 vs. $\bar{g}$ = −0.36, respectively).

Student grade level. We treated grade level as a continuous variable. When controlling for methodological quality, the meta-regression revealed a significant negative association between average student grade level and magnitude of effect for the traditional ES (b = −0.20, SE = 0.06, p < .001). This effect suggests that the science achievement gap between ELL and non-ELL students fades in traditional instruction across higher grade levels. No other moderation effects involving grade level were significant.

Professional development. Whereas the dosage of professional development was not a significant moderator for any of the effects of interest, the focus of professional development training moderated the findings for treatment ES (Q_B = 8.74, df = 2, p = .013). Studies in which professional development focused on supporting ELL students yielded larger treatment ESs ($\bar{g}$ = 0.32) than those that did not report focusing on ELL students' academic needs ($\bar{g}$ = 0.06). Professional development did not moderate the inquiry ES, and there were too few studies to examine its potential moderating effect on traditional ES.

Length of treatment. Although the length of treatment moderated the findings for the treatment ES and traditional ES, the moderation effects were very small. Specifically, we found a significant negative association between the length of treatment (in weeks) and magnitude of effect for the treatment ES (b = −0.01, SE = 0.01, p = .03) and traditional ES (b = −0.02, SE = 0.004, p < .001).

Sensitivity Analysis and Robustness Checks

In an effort to examine the robustness of the overall mean effect size estimates, a series of sensitivity analyses were conducted.

Robust variance estimation. To assess the impact of using alternative statistical methods for handling dependence on our findings, we reanalyzed the data set using robust variance estimation (RVE). Results from the RVE meta-analysis were virtually identical to those produced in the standard meta-analysis across all effect size estimates: treatment ES = +0.31, SE = 0.08, df = 20.9, p < .01; inquiry ES = −0.31, SE = 0.07, df = 28.2, p < .001; traditional ES = −0.46, SE = 0.12, df = 14, p < .01). Taken altogether, these analyses suggest that the effect size estimates reported

TABLE 1
Overall Weighted Mean Treatment Effect Size (ES) for Subgroup Analyses of Categorical Moderators

			Treatment ES and 95% CI				Test of Difference	
Moderator	*n*	*k*	$\bar{g}$	*SE*	Lower	Upper	Q_B	*df*
Publication status							8.19**	1
Published	7,595	17	0.37***	0.07	0.24	0.51		
Unpublished	696	6	−0.04	0.11	−0.26	0.18		
Research design								
Randomized experiment	6,161	15	0.18*	0.07	0.04	0.33	5.20*	1
Quasi-experiment	2,130	8	0.46***	0.10	0.26	0.65		
Measurement design							0.01	1
Pretest and posttest	3,651	13	0.27*	0.11	0.06	0.48		
Posttest only	4,154	10	0.27*	0.09	0.10	0.45		
Assessment format							2.22	2
Multiple choice	6,248	14	0.27***	0.09	0.10	0.44		
Constructed response	460	2	0.58*	0.23	0.13	1.04		
Mixed	1,583	7	0.19	0.12	−0.05	0.43		
Assessment type							5.08*	1
Researcher-developed	2,976	13	0.39***	0.08	0.23	0.55		
Standardized	4,829	10	0.12	0.10	−0.06	0.33		
Professional development							4.08	2
Small dose (14 hours)	1,175	5	0.19	0.13	−0.06	0.44		
Large dose (15+ hours)	6,822	16	0.27***	0.07	0.14	0.40		
Not reported	294	2	0.66***	0.19	0.28	1.04		
Professional development							8.74*	2
Focused on English language learners	7,125	15	0.32***	0.06	0.19	0.44		
Not focused on English language learners	872	6	0.06	0.11	−0.16	0.27		
Not reported	294	2	0.67***	0.17	0.30	1.03		
Student grade level							6.77	5
First	420	2	0.35†	0.21	−0.06	0.76		
Second	220	1	0.38	0.27	−0.15	0.92		
Fourth	601	3	0.63***	0.16	0.32	0.94		
Fifth	5,625	9	0.22**	0.09	0.05	0.40		
Sixth	1,058	5	0.24†	0.12	−0.01	0.48		
Mixed	367	3	0.11	0.17	−0.22	0.44		

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

in the meta-analysis are robust to the statistical approach used to model correlated effect sizes (see Appendix B for results).

Outliers. Two effect sizes were identified as outliers in the treatment ES analyses and substituted with Winsorize-adjusted values. Sensitivity analysis revealed that based on the original sample, the overall treatment ES increases to +0.36 ($SE = 0.09$, $p < .001$). Excluding these outliers from the analysis reduces the overall treatment ES to +0.22 ($SE = 0.05$, $p < .001$), which remains both positive and statistically significant. Thus, our interpretations of the findings remain the same with or without adjustment of the two outliers.

Study quality. The quality of evidence used to estimate the effectiveness of inquiry instruction for ELL students was encouraging as the majority of included studies were rated as strong or moderate. However, sensitivity analysis revealed that low-quality studies were associated with larger effect sizes and therefore may be overestimating the benefits of inquiry instruction for ELL students (see Appendix C for details). To assess whether our findings were driven by low-quality studies, we restricted our sample to studies with overall methodological quality ratings of either strong or moderate. Results from this analysis produced an overall treatment ES of +0.22 ($SE = 0.07$, $p = .003$). Thus, the pattern of results and our substantive interpretations of them

TABLE 2
Overall Weighted Mean Inquiry Effect Size (ES) for Subgroup Analyses of Categorical Moderators

			Inquiry ES and 95% CI				Test of Difference	
Moderator	*n*	*k*	$\bar{g}$	*SE*	Lower	Upper	Q_B	*df*
Publication status							0.90	1
Published	24,383	21	−0.36***	0.08	−0.52	−0.20		
Unpublished	23,776	9	−0.22†	0.12	−0.46	0.02		
Research design							0.41	1
Randomized experiment	43,408	23	−0.34***	0.08	−0.49	−0.19		
Quasi-experiment	4,751	7	−0.23	0.15	−0.52	0.05		
Measurement design							14.52***	1
Pretest and posttest	31,590	20	−0.17*	0.08	−0.31	−0.04		
Posttest only	16,569	10	−0.66	0.11	−0.87	−0.45		
Assessment format							2.29	3
Multiple choice	40,315	18	−0.29***	0.08	−0.46	−0.13		
Constructed response	457	3	−0.67*	0.29	−1.23	−0.11		
Mixed	6,670	8	−0.37**	0.13	−0.61	−0.12		
Other	717	1	−0.05	0.36	−0.76	0.67		
Assessment type							5.03*	1
Researcher-developed	32,923	23	−0.24***	0.07	−0.38	−0.10		
Standardized	15,236	7	−0.56***	0.12	−0.79	−0.32		
Professional development							5.79†	2
Small dose (14 hours)	22,310	11	−0.12	0.11	−0.34	0.10		
Large dose (15+ hours)	24,630	17	−0.46**	0.09	−0.63	−0.28		
Not reported	1,219	2	−0.16	0.26	−0.66	0.34		
Professional development							3.83	2
Focused on English language learners	17,842	15	−0.45***	0.10	−0.64	−0.26		
Not focused on English language learners	29,815	14	−0.19*	0.10	−0.40	0.00		
Not reported	502	1	−0.26	0.35	−0.95	0.42		
Grade level							6.68	4
Third	6,299	2	−0.26	0.25	−0.75	−0.24		
Fourth	11,730	7	−0.20	0.14	−0.48	−0.07		
Fifth	21,652	8	−0.54	0.12	−0.78	−0.30		
Sixth	7,271	6	−0.38	0.15	−0.67	−0.08		
Mixed	1,207	7	−0.08	0.15	−0.37	−0.21		

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

remained the same with or without the inclusion of methodologically weak studies.

Assessment of publication bias. Two statistical analyses provide no evidence supporting the presence of publication bias. Egger's linear regression test was performed to evaluate the extent to which a study's effect is related to its sample size by regressing the effect size estimate of a study against the precision of the study, indexed by its standard error. We found no association between these factors (bias coefficient = −0.83, *SE* = 1.40, $p = .56$). An analysis of the funnel plot shown in Figure 5 did not identify any effect sizes that needed to be trimmed or filled, suggesting that publication bias does not likely impact our findings.

Discussion

Question 1: Is Inquiry Instruction Effective for Teaching Science to ELL Students?

The primary purpose of this meta-analytic review was to examine whether inquiry instruction is an effective approach for teaching science to elementary-grade ELL students relative to traditional approaches of science instruction. Overall, we found that ELL students who received inquiry instruction demonstrated gains in science scores that were approximately one-quarter of a standard deviation higher than their ELL peers receiving traditional, direct instruction. Our findings extend past research on science teaching and learning with mainstream K–12 students, including low-performing and at-risk

TABLE 3
Overall Weighted Mean Traditional Effect Size (ES) for Subgroup Analyses of Categorical Moderators

			Traditional ES and 95% CI				Test of Difference	
Moderator	*n*	*k*	$\bar{g}$	*SE*	Lower	Upper	Q_B	*df*
Publication status							2.39	1
Published	11,986	10	−0.58***	0.14	−0.85	−0.31		
Unpublished	3,093	5	−0.21	0.19	−0.59	0.17		
Research design							1.95	1
Randomized experiment	11,676	8	−.60***	0.15	−0.76	−0.14		
Quasi-experiment	3,403	7	−.29†	0.17	−0.61	0.04		
Measurement design							16.18***	1
Pretest and posttest	7,606	9	−0.24*	0.11	−0.44	−0.03		
Posttest only	7,473	6	−0.92	0.13	−1.18	−0.66		
Assessment format							0.39	2
Multiple choice	11,135	8	−0.55***	0.15	−0.86	−0.25		
Constructed response	310	2	−0.49	0.34	−1.13	0.16		
Mixed	3,634	5	−0.40*	0.32	−0.78	−0.02		
Assessment type							2.92†	1
Researcher-developed	5,320	9	−0.36**	0.13	−0.61	−0.11		
Standardized	9,759	6	−0.69***	0.15	−0.99	−0.40		
Grade level							6.83†	3
Fourth	1,591	3	−0.42†	0.25	−0.91	0.08		
Fifth	9,761	5	−0.76***	0.18	−1.12	−0.40		
Sixth	3,168	4	−0.45*	0.21	−0.87	−0.04		
Mixed	559	3	0.05	0.25	−0.44	0.54		

†$p < .10$. *$p < .05$. **$p < .01$. ***$p < .001$.

TABLE 4
Meta-Regression of Continuous Variables on Overall Weighted Mean Effect Sizes (ES)

Moderator	Treatment ES	Inquiry ES	Traditional ES
Constant	0.613*	−0.582	0.652†
Methodological controls			
Published study	0.433***	−0.018	−0.216
Randomized experiment	−0.204	−0.111	−0.024
Pretest and posttest design	−0.114	0.516***	0.429***
Continuous predictors			
Student grade level	−0.049	0.001	−0.197***
Instruction (weeks)	−0.012*	0.005	−0.016***
Professional development (hours)	−0.001	−0.041	—
Number of studies (*k*)	21	27	15
Between-study variance (τ^2)	0.01	0.05	0.01
Heterogeneity (I^2), %	57	92	65

Note. Random effects models were used in all meta-regression analyses. Random effects variance components were estimated using maximum likelihood. Effect sizes computed as Hedge's *g*. Reference group for controls = unpublished study, quasi-experiment, posttest-only design.
†$p < .10$. *$p < .05$. ***$p < .001$.

non-ELL students (Hill, Bloom, Black, & Lipsey, 2008; Lipsey et al., 2012) to young, elementary school aged ELL students.

Despite the pedagogical and theoretical concerns about inquiry instruction for ELL students (i.e., Kirschner et al., 2006; Secker, 2002; Tobias & Duffy, 2009), most studies found that inquiry instruction was either as effective or more effective than traditional science instruction for ELL students. There was only one of the studies we identified

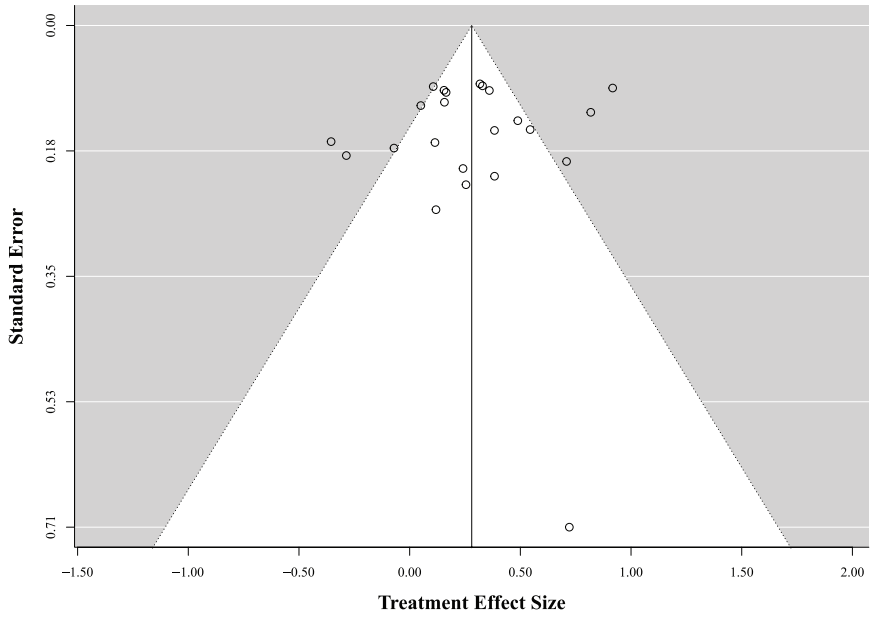

FIGURE 5. *Funnel plot used for evaluating publication bias. Average weighted treatment effect size estimated for each study (horizontal axis) plotted against corresponding standard error (vertical axis).*

reporting that inquiry instruction was significantly worse for ELL students than traditional instruction. Further, the 95% prediction interval ranged from −0.30 to 0.86. Assuming these true effect sizes are normally distributed about the mean (Borenstein et al., 2009), we can predict that about 71% of future studies would yield a meaningful positive effect (between 0.10 and 0.86) favoring inquiry instruction, and 7% of future studies would yield a meaningful negative effect (between −0.10 and −0.30) favoring traditional instruction. Thus, our findings mitigate concerns that inquiry instruction may hinder science outcomes for ELL students.

While these findings show that ELL students have stronger science outcomes from inquiry instruction compared to traditional approaches to science instruction, further research is needed to better understand why. Although several compelling reasons have been suggested in the literature, such as rich peer-to-peer collaboration and hands-on learning, perhaps the most widely cited explanation is that inquiry instruction reduces the reliance on language for engaging with and understanding scientific content through hands-on learning activities that emphasize nonverbal forms of processing and participation (Huerta & Jackson, 2010; Lee & Buxton, 2013). The active learning involved in inquiry instruction is thought to maximize meaningful learning opportunities for ELL students by diminishing the heavy linguistic demands associated with traditional forms of textbook and lecture-based learning (Fang, 2006; Lewis, Lee, Santau, & Cone, 2010). These practices align with instructional strategies considered effective for ELL students (Echevarria et al., 2011; Goldenberg, 2013). Because the advantages were greater when teachers received professional development focused on supporting ELL students, it is possible that inquiry instruction's benefits may stem from its alignment with best practices for ELL students. However, further research is needed to explore the mechanism by which inquiry instruction provides stronger learning outcomes for ELL students.

Question 2: Does Inquiry Instruction Provide Comparable Effects for ELL and Non-ELLs?

In addition to examining the effectiveness of inquiry instruction for ELL students, we explored whether ELL students experienced learning benefits that are comparable to those enjoyed by their non-ELL peers. Although non-ELL students showed stronger science outcomes using both instructional models, the achievement gap between ELL and non-ELL students was diminished with inquiry instruction. Whereas the achievement may be diminished further by restricting our findings to studies that used pretest-posttest designs, the more limited gains shown by ELL students remain statistically significant and practically important. ELL students' learning outcomes may have been adversely affected by their limited proficiency in English and weak mastery of the academic language in ways that non-ELL students are not affected (Lee, 2005). For example, common linguistic features in science texts and discourse that may interfere with ELL students' comprehension and learning include the frequent use of ordinary words with nonvernacular meanings, complex sentence structures, and even passive voice (Fang, 2006). In contrast, non-ELL students are relatively unaffected by these factors and therefore may be better able to construct scientific knowledge (Bresser & Fargason,

2013; Fang, 2006; Janzen, 2008; Mayer, 2004). Thus, while inquiry science instruction may better support learning for ELL students than traditional instructions by mitigating the linguistic demands through the use of hands-on activities, it may not be enough to fully remedy the comprehension difficulties ELL students experience in science class (August et al., 2009; Bravo & Cervetti, 2014; Kieffer et al., 2009).

Alternately, the differentially smaller effects of inquiry instruction on ELL students' science achievement may result from their more limited prior experience with science. There have been concerns that although inquiry instruction may provide rich benefits to high-performing students by providing them with opportunities to apply, test, and deepen their understanding of science (Kirschner et al., 2006; Klahr & Nigam, 2004), students with less exposure to science are less likely to have the necessary schemata in place to construct deep meaning through the self-guided exploration. Consequently, these students are less likely to develop the same depth of understanding of the scientific principles they are exploring (Kendeou & van den Broek, 2007; Norman & Schmidt, 1992). Thus, despite showing greater achievement through inquiry instruction than traditional science instruction, ELL students may not have sufficient background knowledge to reap the same benefits from inquiry instruction as their non-ELL peers.

In sum, although the evidence suggests that inquiry instruction may differentially benefit ELL and non-ELL students, it did, however, promote substantive gains in science for both groups of students. Although preliminary, findings from this review also suggest that inquiry instruction may help diminish the achievement gap between ELL and non-ELL students in science. Based on this evidence, we argue that inquiry instruction has the potential to effectively educate a diverse body of students and recommend that further efforts be taken to understand how inquiry instruction can be adapted to better serve ELL students' instructional needs and further reduce the achievement gap between ELL and non-ELL students in science.

Question 3: What Factors Moderate the Effect of Inquiry Instruction for ELLs?

The final aim of this study was to investigate factors that may moderate the effectiveness of inquiry instruction for ELL students. One such factor was professional development. We found equivocal evidence concerning the effects of the duration of professional development. Although the moderation analysis found that the duration of professional development was unrelated to the impact of inquiry instruction on ELL students' science achievement, partitioning our sample into subgroups with at least 15 hours or less than 15 hours of professional development provided suggestive results consistent with Yoon et al. (2007). More specifically, we found that studies with at least 15 hours of professional development yielded positive and statistically significant treatment effects whereas those that offered less professional development time did not. In contrast, we found the focus, or content, of professional development yielded a far more robust effect. Studies in which professional development focused on building teachers' skills and knowledge to address ELL students' academic and linguistics needs yielded greater positive effect sizes than those that did not. Taken together, our findings add to a growing body of work suggesting that a minimal threshold of professional development training geared toward the needs of the target student population may better enable teachers to develop the pedagogical skills and knowledge to implement inquiry instruction in a way that maximizes ELL student learning (Darling-Hammond, Chung-Wei, Andree, Richardson, & Orphanos, 2009; Garet, Porter, Desimone, Birman, & Yoon, 2001).

Methodological features, such as the study's publication status, research design, and assessment type, also moderated the reported effects of inquiry instruction on ELL students' science achievement. In contrast to what is commonly found in meta-analytic research (Lazonder & Harmsen, 2016; Schroeder, Scott, Tolson, Huang, & Lee, 2007; Seidel & Shavelson, 2007), published studies and quasi-experiments produced significantly lower effect sizes than unpublished studies and randomized experiments. These findings suggest that differences in methodological quality are correlated with effect size estimates. Furthermore, researcher-developed assessments yielded significantly greater effect sizes than standardized assessment, which may be attributed to the proximal versus distal nature of the assessment used (Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). That is, researcher-developed assessments tend to be more closely aligned with the content of the intervention than standardized assessments. Thus, researcher-developed assessments tend to be more sensitive to the impact of instructional intervention on student achievement, resulting in larger effect sizes.

Alternately, this effect may reflect the degree to which assessments present ELL students with the opportunity to demonstrate their science knowledge. In developing an assessment instrument, standardized tests are restricted almost exclusively to multiple choice items, whereas researcher-developed assessments tend to be more amenable to open-ended and constructed response items. Constructed response items may offer a more meaningful measure of achievement insofar as they enable students, particularly ELL students, to express their scientific understanding using their own words—without being constrained by the linguistically demanding and potentially confusing options provided in multiple-choice formats (Turkan & Liu, 2012). Our meta-analytic findings provide some evidence for this assumption as effect sizes tended to be higher for constructed response items than

multiple-choice items. However, the limited number of studies that reported outcomes separately for constructed response and multiple-choice item formats tempers our confidence in these results. Indeed, further research is needed to disentangle whether one method of assessment offers more valid inferences than another.

Research Limitations and Future Directions

Although our findings contribute to the literature, they should be interpreted in light of the study's limitations. First, the stringent study selection criteria applied in the study selection phase led to the exclusion of a large number of qualitative studies that did not provide the statistical information needed to calculate effect sizes. Although these studies may provide valuable descriptive insight into the experiences of ELL students during inquiry instruction, they did not report the data needed to conduct a quantitative synthesis. We recommend that future research consult qualitative studies for evidence of additional contextual factors that may explain variability in the effectiveness of inquiry-based instruction.

Second, rather than include the full range of English proficiency of language minority students in this meta-analysis, we focused on current ELL students and excluded former ELL students who were either reclassified as fluent English proficient, exited from English language development programs, or no longer received English language support services. This decision may restrict the generalizability of the findings to a narrow subpopulation of ELL students. However, including the scores of former ELL students, who typically resemble non-ELL students in terms of English proficiency and academic achievement, could obscure the effects of inquiry instruction for current ELL students. Indeed, including former ELL students may risk results that overstate the benefits of inquiry learning for ELL students and understate the science achievement gap between ELL and non-ELL students (cf. Saunders & Marcelletti, 2013). Nonetheless, we recognize this approach as a limitation and therefore recommend that future research focus on investigating the effect of inquiry learning across various levels of English language proficiency.

Third, while the sample size included in this meta-analysis was large enough to compute reliable main effect sizes, we often lacked sufficient data to compute potentially interesting moderating effects, mainly because primary studies failed to report such information. For example, few studies reported salient features of the instructional approach (e.g., time on task) and demographics information pertaining to students (e.g., race/ethnicity, gender, primary language spoken at home) and teachers (e.g., years of teaching experience, educational attainment, pedagogical training), which have important theoretical and pedagogical implications. As such, we recommend that future research in science teaching include clear descriptions of not only the instructional approach but also their student and teacher samples.

Conclusion

This synthesis contributes to the field and advances our current understanding of evidence-based science pedagogy in several ways. For example, despite a growing body of individual studies suggesting that inquiry instruction is a particularly effective approach for teaching science ELL students, conflicting results and a lack of empirical consensus have rendered the precise nature and magnitude of its effects unclear. In an effort to address this gap, our study systematically surveyed the literature on inquiry instruction and synthesized empirical findings from over a decade's worth of research. Consequently, results from this study can be used to inform national dialogue concerning effective, appropriate, and equitable instructional practices for linguistically diverse students, which is essential given the growing number of ELL students in U.S.-based public schools. These findings have never been more important as educational practitioners, researchers, and policymakers deliberate solutions to pressing issues facing K–12 schooling and make critical decisions that will undoubtedly shape the future of science education for all students. Finally, this study serves as a comprehensive synthesis of what the current state of literature reveals about the effectiveness of inquiry instruction for ELL students, providing valuable information and insight for those interested in advancing the frontier in opportune ways. Such guidance is particularly important given the extraordinary amount of resources involved in the planning and execution of truly effective and impactful research.

Overall, the findings from our work suggest that inquiry instruction has the potential to improve science learning and performance for not only English-proficient students but also ELL students—albeit to a lesser extent. Although the learning benefits associated with inquiry instruction are compelling, our data suggest that ELL students might require additional academic and linguistic support if they are to attain a level of science achievement that is on par with their English-proficient peers. Therefore, we urge researchers to not only continue to explore the nature of inquiry instruction through applied experimental investigation but also to investigate other avenues for improving ELL students' science learning and performance, in particular, research programs aimed at investigating the mechanisms by which inquiry-based instruction leads to improved achievement, how technology can be used to support such outcomes, and whether the NGSS provides other opportunities for ELL students to access quality science education.

APPENDIX A

Study Characteristics and Key Moderators for Studies Included in the Meta-Analysis

Study by Authors	Publication Status	Research Design	Measurement Design	Grade Level	Treatment Effect Size	Inquiry Effect Size	Traditional Effect Size
Amaral, Garrison, and Klentschy (2002)[a]	Published	Quasi-experimental	Posttest only	Fourth	0.25	−0.64	−0.27
Amaral et al. (2002)[b]	Published	Quasi-experimental	Posttest only	Sixth	0.82	−0.59	−0.71
August, Branum-Martin, Cardenas-Hagan, and Francis (2009)	Published	Experimental	Pre-posttest	Fifth	0.33	−0.25	−0.18
August et al. (2014)	Published	Experimental	Pre-posttest	Sixth	0.16	−0.80	−0.66
Bravo and Cervetti (2014)	Published	Experimental	Pre-posttest	Fourth, fifth	0.71	−0.04	0.06
Cervetti, Kulikowich, and Bravo (2015)[a]	Published	Experimental	Pre-posttest	Fourth, fifth	—	−0.06	—
Cervetti et al. (2015)[b]	Published	Experimental	Pre-posttest	Fourth, fifth	—	0.04	—
Diamond, Maerten-Rivera, Rohrer, and Lee (2014)	Published	Experimental	Posttest only	Fifth	—	−0.22	—
Fradd, Lee, Sutman, and Saxton (2001)[a]	Published	Experimental	Pre-posttest	Fourth	0.92	−0.26	—
Fradd et al. (2001)[b]	Published	Experimental	Pre-posttest	Fourth	0.92	—	—
Fulp, Warren, and Banilower (2009)[a]	Published	Experimental	Pre-posttest	Third	—	−0.12	—
Fulp et al. (2009)[b]	Published	Experimental	Pre-posttest	Fourth	—	−0.02	—
Fulp et al. (2009)[c]	Published	Experimental	Pre-posttest	Fifth	—	−0.04	—
Fulp et al. (2009)[d]	Published	Experimental	Pre-posttest	Sixth	—	−0.09	—
Heller, Daehler, Wong, Shinohara, and Miratrix (2012)[a]	Published	Experimental	Pre-posttest	Fourth	0.54	—	−0.29
Heller et al. (2012)[b]	Published	Experimental	Pre-posttest	Fourth	—	−0.12	—
Heller et al. (2012)[c]	Published	Experimental	Pre-posttest	Fourth	—	−0.23	—
Heller et al. (2012)[d]	Published	Experimental	Pre-posttest	Fourth	—	−0.21	—
Huerta, Irby, Lara-Alecio, and Tong (2015)	Published	Experimental	Posttest only	Fifth, sixth	—	−0.83	—
Lara-Alecio et al. (2012)	Published	Experimental	Posttest only	Fifth	0.24	—	—
Lee et al. (2004–2007)	Published	Experimental	Pre-posttest	Fifth, sixth	—	−0.38	—
Llosa et al. (2016)	Published	Experimental	Pre-Posttest	Fifth	0.16	−0.12	−0.34
Lymer (2013)	Unpublished	Quasi-experimental	Posttest only	First	0.12	—	—
Maerten-Rivera, Ahn, Lanier, Diaz, and Lee (2016)[a]	Published	Experimental	Posttest only	Fifth	0.05	−1.14	−1.12
Maerten-Rivera et al. (2016)[b]	Published	Experimental	Posttest only	Fifth	0.16	−0.92	−1.00
Maerten-Rivera et al. (2016)[c]	Published	Experimental	Posttest only	Fifth	0.32	−0.93	−1.14
Miller, Jaciw, and Ma (2007)	Unpublished	Experimental	Pre-Posttest	Third, fifth	−0.07	−0.13	0.20
Miller, Jaciw, and Vu (2007)	Unpublished	Experimental	Pre-Posttest	Third, fifth	−0.29	−0.06	−0.11
Pyke, Lynch, Kuipers, Szesze, and Watson (2004)	Unpublished	Experimental	Pre-posttest	Sixth	0.11	−0.34	−0.17
Pyke, Lynch, Kuipers, Szesze, and Watson (2005)	Unpublished	Experimental	Pre-posttest	Sixth	−0.35	−0.23	0.05
Pyke, Lynch, Kuipers, Szesze, and Watson (2006)	Unpublished	Experimental	Posttest only	Sixth	0.38	−0.95	−1.06
Santau, Maerten-Rivera, and Huggins (2011)	Published	Experimental	Pre-posttest	Fourth	—	−0.04	—
Shaw and Nagashima (2009)	Published	Experimental	Posttest only	Fifth	—	−0.05	—
Shaw, Lyon, Stoddart, Mosqueda, and Menon (2014)	Published	Experimental	Pre-posttest	Third, sixth	—	0.07	—

(continued)

APPENDIX A (CONTINUED)

Study by Authors	Publication Status	Research Design	Measurement Design	Grade Level	Treatment Effect Size	Inquiry Effect Size	Traditional Effect Size
Tong, Irby, Lara-Alecio, and Koch (2014)	Published	Experimental	Posttest only	Fifth	0.72	—	—
Zwiep and Straits (2013)[a]	Published	Quasi-experimental	Pre-posttest	Fifth	0.11	—	—
Zwiep and Straits (2013)[b]	Published	Quasi-experimental	Pre-posttest	Fifth	0.36	—	—
Zwiep and Straits (2013)[c]	Published	Quasi-experimental	Pre-posttest	First	0.49	—	—
Zwiep and Straits (2013)[d]	Published	Quasi-experimental	Pre-posttest	Second	0.38	—	—

Note. Studies with alphabetic superscripts refer to multiple independent effect sizes generated from the same study. The data used to calculate an overall effect size for Lee et al. (2004–2007) is based on information reported in Lee, Maerten-Rivera, Penfield, Leroy, and Secada (2008); Lee, Mahotiere, Salinas, Penfield, and Maerten-Rivera (2009); and Lee, Penfield, and Maerten-Rivera (2009).

APPENDIX B

Comparison of Mean Effect Sizes (ES) From Standard and Robust Variance Estimation Meta-Analyses

	Treatment ES		Inquiry ES		Traditional ES	
Statistics	Standard	RVE	Standard	RVE	Standard	RVE
Hedge's *g* effect size	0.28	0.31	−0.31	−0.31	−0.46	−0.46
Standard error	0.07	0.08	0.07	0.07	0.12	0.12
95% CI low estimate	0.15	0.13	−0.45	−0.44	−0.70	−0.71
95% CI high estimate	0.41	0.48	−0.18	−0.18	−0.22	−0.20
Degrees of freedom (*df*)	—	20.9	—	28.2	—	14
Heterogeneity (I^2), %	82	85	92	93	95	96
Between-study variance (τ^2)	0.07	0.10	0.12	0.13	0.21	0.25
Number of studies (*k*)	23	45	30	59	15	20
Correlation (ρ)	—	0.80	—	0.80	—	0.80

Note. To achieve independence, standard meta-analyses were conducted using synthetic effect sizes, whereas RVE meta-analyses used correlated effects models with small-sample bias corrections. RVE = robust variance estimation.

APPENDIX C

Examining the Effectiveness of Inquiry Instruction for English Language Learner Students by Overall Study Quality Ratings

		Treatment Effect Size and 95% Confidence Interval						Test of Difference		
Quality Rating	*k*	$\bar{g}$	*SE*	Lower	Upper	*p*	I^2 (%)	Q_B	*df*	*p*
Individual studies								3.11	2	.21
Strong	6	0.20	0.11	−0.02	0.42	0.074	21			
Moderate	11	0.23	0.10	0.04	0.42	0.020	76			
Weak	6	0.46	0.12	0.23	0.69	0.000	91			
Combined studies										
High quality	17	0.22	0.07	0.07	0.35	0.003	66	3.27	1	.07
Low quality	6	0.46	0.12	0.23	0.69	0.000	91			

Note. Appraisal of study quality measured using the Quality Assessment Tool for Quantitative Studies. High-quality group composed of studies with overall quality ratings of strong and moderate. Low-quality group composed of studies with overall quality ratings of weak.

Authors' Note

GE and JA are each supported by the National Science Foundation (Grants No. DGE-1321846), and SMJ is supported by the National Institute on Aging (Grant No.1K02AG054665-01). SMJ has an indirect financial interest in the MIND Research Institute, whose interests are related to this work.

References

*References marked with an asterisk indicate studies included in the meta-analysis.

Achieve. (2012). *Next Generation Science Standards*. Retrieved from http://www.nextgenscience.org/next-generation-science-standards

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1–18. doi:10.1037/a0021017

*Amaral, O. M., Garrison, L., & Klentschy, M. (2002). Helping English learners increase achievement through inquiry-based science instruction. *Bilingual Research Journal*, *26*(2), 213–239. doi:10.1080/15235882.2002.10668709

*August, D., Branum-Martin, L., Cardenas-Hagan, E., & Francis, D. J. (2009). The impact of an instructional intervention on the science and language learning of middle grade English language learners. *Journal of Research on Educational Effectiveness*, *2*(4), 345–376. doi:10.1080/19345740903217623

*August, D., Branum-Martin, L., Cárdenas-Hagan, E., Francis, D. J., Powell, J., Moore, S., & Haynes, E. F. (2014). Helping ELLs meet the Common Core State Standards for literacy in science: The impact of an instructional intervention focused on academic language. *Journal of Research on Educational Effectiveness*, *7*(1), 54–82. doi:10.1080/19345747.2013.836763

Baker, D. P., Fabrega, R., Galindo, C., & Mishook, J. (2004). Instructional time and national achievement: Cross-national evidence. *Prospects*, *34*(3), 311–334. doi:10.1007/s11125-004-5310-1

Becker, B. J., & Wu, M. J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science*, *22*(3), 414–429.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. West Sussex: John Wiley & Sons, Ltd.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2014). *Comprehensive meta-analysis version 3* [Computer software]. Englewood, NJ: Biostat.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness*, *1*(4), 237–264. doi:10.1080/19345740802328273

*Bravo, M. A., & Cervetti, G. N. (2014). Attending to the language and literacy needs of English learners in science. *Equity & Excellence in Education*, *47*(2), 230–245. doi:10.1080/10665684.2014.900418

Bresser, R., & Fargason, S. (2013). *Becoming scientists: Inquiry-based teaching in diverse classrooms, Grades 3–5*. Portland, ME: Stenhouse Publishers.

Bruner, J. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

*Cervetti, G. N., Kulikowich, J. M., & Bravo, M. A. (2015). The effects of educative curriculum materials on teachers' use of instructional strategies for English language learners in science and on student learning. *Contemporary Educational Psychology*, *40*, 86–98. doi:10.1016/j.cedpsych.2014.10.005

Cuevas, P., Lee, O., Hart, J., & Deaktor, R. (2005). Improving science inquiry with elementary students of diverse backgrounds. *Journal of Research in Science Teaching*, *42*(3), 337–357. doi:10.1002/tea.20053

Darling-Hammond, L., Chung-Wei, R., Andree, A., Richardson, N., & Orphanos, S. (2009). *Professional learning in the learning profession: A status report on teacher development in the U.S. and abroad*. Dallas, TX: National Staff Development Council.

*Diamond, B. S., Maerten-Rivera, J., Rohrer, R. E., & Lee, O. (2014). Effectiveness of a curricular and professional development intervention at improving elementary teachers' science content knowledge and student achievement outcomes: Year 1 results. *Journal of Research in Science Teaching*, *51*(5), 635–658.

Dorph, R., Shields, P., Tiffany-Morales, J., Hartry, A., & McCaffrey, T. (2011). *High hopes-few opportunities: The status of elementary science education in California*. Sacramento, CA: The Center for the Future of Teaching and Learning at WestEd.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. doi:10.1111/j.0006-341X.2000.00455.x

Echevarria, J., Richards-Tutor, C., Canges, R., & Francis, D. (2011). Using the SIOP model to promote the acquisition of language and science concepts with English learners. *Bilingual Research Journal*, *34*(3), 334–351. doi:10.1080/15235882.2011.623600

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. doi:10.1136/bmj.315.7109.629

Fang, Z. (2006). The language demands of science reading in middle school. *International Journal of Science Education*, *28*(5), 491–520. doi:10.1080/09500690500339092

Fisher, Z., & Tipton, E. (2014). *robumeta: An R-package for robust variance estimation in meta-analysis*. Retrieved from https://arxiv.org/pdf/1503.02220.pdf

*Fradd, S. H., Lee, O., Sutman, F. X., & Saxton, M. K. (2001). Promoting science literacy with English language learners through instructional materials development: A case study. *Bilingual Research Journal*, *25*(4), 479–501. doi:10.1080/15235882.2001.11074464

*Fulp, S. L., Warren, C. L., & Banilower, E. R. (2009). *Science: It's elementary. Year three evaluation report*. San Jose, CA: Horizon Research, Inc.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, *82*(3), 300–329. doi:10.3102/0034654312457206

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*, 915–945. doi:10.3102/00028312038004915

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English language learners in US schools: An overview of research findings. *Journal of Education for Students Placed at Risk*, *10*(4), 363–385.

Goldenberg, C. (2013). Unlocking the research on English learners: What we know and don't yet know about effective instruction. *American Educator*, *37*(2), 4–13.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*(2), 490.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65.

*Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333–362. doi:10.1002/tea.21004

Higgins, J. P. T., Altman, D. G., & Sterne, J. A. C. (2017). Chapter 8: Assessing risk of bias in included studies. In J. P. T. Higgins, R. Churchill, J. Chandler, & M. S. Cumpston (Eds.), *Cochrane handbook for systematic reviews of interventions version 5.2.0*. Retrived from www.training.cochrane.org/handbook.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, *327*(7414), 557. doi:10.1136/bmj.327.7414.557

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. doi:10.1111/j.1750-8606.2008.00061.x

*Huerta, M., Irby, B. J., Lara-Alecio, R., & Tong, F. (2015). Relationship between language and concept science notebook scores of English language learners and/or economically disadvantaged students. *International Journal of Science and Mathematics Education*, *14*(2), 269–285. doi:10.1007/s10763-015-9640-7

Huerta, M., & Jackson, J. (2010). Connecting literacy and science to increase achievement for English language learners. *Early Childhood Education Journal*, *38*(3), 205–211. doi:10.1007/s10643-010-0402-4

Huerta, M., Tong, F., Irby, B. J., & Lara-Alecio, R. (2016). Measuring and comparing academic language development and conceptual understanding via science notebooks. *The Journal of Educational Research*, *109*(5), 503–517. doi:10.1080/00220671.2014.992582

Janzen, J. (2008). Teaching English language learners in the content areas. *Review of Educational Research*, *78*(4), 1010–1038. doi:10.3102/0034654308325580

Jiang, F., & McComas, W. F. (2015). The effects of inquiry teaching on student science achievement and attitudes: Evidence from propensity score analysis of PISA data. *International Journal of Science Education*, *37*(3), 554–576. doi:10.1080/09500693.2014.1000426

Kendeou, P., & van den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition*, *35*(7), 1567–1577. doi:/10.3758/BF03193491

Kieffer, M. J., Lesaux, N., Rivera, M., & Francis, D. J. (2009). Accommodations for English language learners taking large scale assessments: A meta-analysis on effectiveness and validity. *Review of Educational Research*, *29*(3), 1168–1201. doi:10.3102/0034654309332490

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. doi:10.1207/s15326985ep4102_1

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667. doi:10.1111/j.0956-7976.2004.00737.x

*Lara-Alecio, R., Tong, F., Irby, B. J., Guerrero, C., Huerta, M., & Fan, Y. (2012). The effect of an instructional intervention on middle school English learners' science and English reading achievement. *Journal of Research in Science Teaching*, *49*(8), 987–1011. doi:10.1002/tea.21031

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. *Review of Educational Research*, *86*(3), 681–718. doi:10.3102/0034654315627366

Lee, O. (2005). Science education and English language learners: Synthesis and research agenda. *Review of Educational Research*, *75*(4), 491–530. doi:10.3102/00346543075004491

Lee, O., & Buxton, C. A. (2013). Teacher professional development to improve science and literacy achievement of English language learners. *Theory Into Practice*, *52*(2), 110–117. doi:10.1080/00405841.2013.770328

Lee, O., Deaktor, R., Enders, C., & Lambert, J. (2008). Impact of a multiyear professional development intervention on science achievement of culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching*, *45*(6), 726–747. doi:10.1002/tea.20231

Lee, O., Deaktor, R., Hart, J. E., Cuevas, P., & Enders, C. (2005). An instructional intervention's impact on the science and literacy achievement of culturally and linguistically diverse elementary students. *Journal of Research in Science Teaching*, *42*(8), 857–887. doi:10.1002/tea.20071

*Lee, O., Maerten-Rivera, J., Penfield, R. D., LeRoy, K., & Secada, W. G. (2008). Science achievement of English language learners in urban elementary schools: Results of a first-year professional development intervention. *Journal of Research in Science Teaching*, *45*(1), 31–52. doi:10.1002/tea.20209

*Lee, O., Mahotiere, M., Salinas, A., Penfield, R. D., & Maerten-Rivera, J. (2009). Science writing achievement among English language learners: Results of three-year intervention in urban elementary schools. *Bilingual Research Journal*, *32*(2), 153–167. doi:10.1080/15235880903170009

*Lee, O., Penfield, R., & Maerten-Rivera, J. (2009). Effects of fidelity of implementation on science achievement gains among English language learners. J*ournal of Research in Science Teaching*, *46*(7), 836–859. doi:10.1002/tea.20335

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English language arts and mathematics. *Educational Researcher*, *42*(4), 223–233. doi:10.3102/0013189X13480524

Lewis, S., Lee, O., Santau, A. O., & Cone, N. (2010). Student initiatives in urban elementary science classrooms. *School Science*

and Mathematics, *110*(3), 160–172. doi:10.1111/j.1949-8594.2010.00018.x

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. Washington, DC: National Center for Special Education Research.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage Publications.

*Llosa, L., Lee, O., Jiang, F., Haas, A., O'Connor, C., Van Booven, C. D., & Kieffer, M. J. (2016). Impact of a large-scale science intervention focused on English language Learners. *American Educational Research Journal*, *53*(2), 395–424. doi:10.3102/0002831216637348

*Lymer, D. (2013). *How does inquiry-based learning affect attitudes towards science of first grade English language learners?* (Unpublished master's thesis). Montana State University, Bozeman, MT.

*Maerten-Rivera, J., Ahn, S., Lanier, K., Diaz, J., & Lee, O. (2016). Science achievement over a three-year curricular and professional development intervention with English language learners in urban elementary schools. *The Elementary School Journal*, *116*(4), 600–624.

Maerten-Rivera, J., Myers, N., Lee, O., & Penfield, R. (2010). Student and school predictors of high-stakes assessment in science. *Science Education*, *94*(6), 937–962. doi:10.1002/sce.20408

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist*, *59*, 14–19.

Meyer, L. M. (2000). Barriers to meaningful instruction for English learners. *Theory Into Practice*, *39*(4), 228–236. doi:10.1207/s15430421tip3904_6

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

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

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—What is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching*, *47*(4), 474–496. doi:10.1002/tea.20347

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. doi:10.1177/1094428106291059

National Center for Education Statistics. (2014). *The condition of education 2014* (NCES 2014-083). Washington, DC: U.S. Department of Education.

National Collaborating Centre for Methods and Tools. (2008). *Quality assessment tool for quantitative studies*. Retrieved from http://www.nccmt.ca/resources/search/14.

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

Norman, G. R., & Schmidt, H. G. (1992). The psychological basis of problem-based learning: A review of the evidence. *Academic Medicine*, *67*(9), 557–565. doi:10.1097/00001888-199209000-00002

*Pyke, C., Lynch, S., Kuipers, J., Szesze, M., & Watson, W. (2004). *Implementation study of Exploring Motion and Forces (2004–2005): SCALE-uP Report No. 8*. Washington, DC: The George Washington University.

*Pyke, C., Lynch, S., Kuipers, J., Szesze, M., & Watson, W. (2005). *Implementation study of Exploring Motion and Forces (2004–2005): SCALE-uP Report No. 8*. Washington, DC: The George Washington University.

*Pyke, C., Lynch, S., Kuipers, J., Szesze, M., & Watson, W. (2006). *Implementation study of Exploring Motion and Forces (2005–2006): SCALE-uP Report No. 13*. Washington, DC: The George Washington University.

Robinson-Cimpian, J. P., Fiske, S. T., Thompson, K. D., & Umansky, I. M. (2016). Research and policy considerations for English learner equity. *Policy Insights From the Behavioral and Brain Sciences*, *3*(1), 129–137. doi:10.1177/2372732215623553

Rosebery, A. S., & Warren, B. (Eds.). (2008). *Teaching science to English language learners: Building on students' strengths*. Arlington, VA: National Science Teachers Association.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. doi:10.1037/0033-2909.86.3.638

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369–393. doi:10.1002/tea.10027

*Santau, A. O., Maerten-Rivera, J. L., & Huggins, A. C. (2011). Science achievement of English language learners in urban elementary schools: Fourth-grade student achievement results from a professional development intervention. *Science Education*, *95*(5), 771–793. doi:10.1002/sce.20443

Saunders, W. M., & Marcelletti, D. J. (2013). The gap that can't go away: The catch-22 of reclassification in monitoring the progress of English learners. *Educational Evaluation and Policy Analysis*, *35*(2), 139–156. doi:10.3102/0162373712461849

Schroeder, C. M., Scott, T. P., Tolson, H., Huang, T. Y., & Lee, Y. H. (2007). A meta-analysis of national research: Effects of teaching strategies on student achievement in science in the United States. *Journal of Research in Science Teaching*, *44*(10), 1436–1460. doi:10.1002/tea.20212

Secker, C. V. (2002). Effects of inquiry-based teacher practices on science excellence and equity. *The Journal of Educational Research*, *95*(3), 151–160. doi:10.1080/00220670209596585

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, *77*(4), 454–499. doi:10.3102/0034654307310317

*Shaw, J. M., Lyon, E. G., Stoddart, T., Mosqueda, E., & Menon, P. (2014). Improving science and literacy learning for English language learners: Evidence from a pre-service teacher preparation intervention. *Journal of Science Teacher Education*, *25*(5), 621–643. doi:10.1007/s10972-013-9376-6

*Shaw, J. M., & Nagashima, S. O. (2009). The achievement of student subgroups on science performance assessments in inquiry-based classrooms. *Electronic Journal of Science Education*, *13*(2), 6–29.

Song, F., Hooper, L., & Loke, Y. (2013). Publication bias: What is it? How do we measure it? How do we avoid it? *Open Access Journal of Clinical Trials*, *2013*(5), 71–81. doi:10.2147/OAJCT.S34419

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13–30. doi:10.1002/jrsm.1091

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. doi:10.1037/met0000011

Tobias, S., & Duffy, T. M. (Eds.). (2009). *Constructivist instruction: Success or failure?* New York, NY: Routledge.

Tolbert, S., Stoddart, T., Lyon, E. G., & Solis, J. L. (2014). The Next Generation Science Standards, Common Core State Standards, and English learners: Using the SSTELLA framework to prepare secondary science teachers. *Issues in Teacher Education*, *23*(1), 65–90.

*Tong, F., Irby, B. J., Lara-Alecio, R., & Koch, J. (2014). Integrating literacy and science for English language learners: from learning-to-read to reading-to-learn. *The Journal of Educational Research*, *107*(5), 410–426. doi:10.1080/00220671.2013.833072

Turkan, S., & Liu, O. L. (2012). Differential performance by English language learners on an inquiry-based science assessment. *International Journal of Science Education*, *34*(15), 2343–2369. doi:10.1080/09500693.2012.705046

Weinburgh, M., Silva, C., Smith, K. H., Groulx, J., & Nettles, J. (2014). The intersection of inquiry-based science and language: Preparing teachers for ELL classrooms. *Journal of Science Teacher Education*, *25*(5), 519–541. doi:10.1007/s10972-014-9389-9

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. (2007). *Reviewing the evidence on how teacher professional development affects student achievement* (Issues and Answers Report, REL 2007 No. 033). Washington, DC: U.S. Department of Education, Regional Educational Laboratory Southwest.

*Zwiep, S. G., & Straits, W. J. (2013). Inquiry science: The gateway to English language proficiency. *Journal of Science Teacher Education*, *24*(8), 13. doi:10.1007/s10972-013-9357-9

Authors

GABRIEL ESTRELLA is a PhD candidate in the School of Education at the University of California, Irvine. His research focuses on evidence-based learning strategies and teaching methods, the academic achievement and motivational outcomes of underrepresented students in science, and meta-analytic research methods.

JACKY AU is a PhD candidate in the Department of Cognitive Science at the University of California, Irvine. He is a meta-analyst, specializing in the field of cognitive training and plasticity.

SUSANNE M. JAEGGI is a faculty member at the School of Education and the Department of Cognitive Sciences at the University of California, Irvine. She studies learning and individual differences across the life span, focusing on the development of cognitive interventions and the investigation of whether and how those interventions generalize to nontrained cognitive domains.

PENELOPE COLLINS is an associate professor in the School of Education at the University of California, Irvine. Her research examines the development of language, literacy, and academic skills for children from linguistically diverse backgrounds and effective educational practices to support English language learners.