*AERA Open*
*July-September 2018, Vol. 4, No. 3, pp. 1–19*
*DOI: 10.1177/2332858418791991*
*© The Author(s) 2018. http://journals.sagepub.com/home/ero*

# Investigating Science Education Effect Sizes: Implications for Power Analyses and Programmatic Decisions

**Joseph A. Taylor**
**Susan M. Kowalski**

*BSCS Science Learning*

**Joshua R. Polanin**

*American Institutes for Research*

**Karen Askinas**
**Molly A. M. Stuhlsatz**
**Christopher D. Wilson**

*BSCS Science Learning*

**Elizabeth Tipton**

*Columbia University*

**Sandra Jo Wilson**

*Abt Associates, Inc.*

*A priori power analyses allow researchers to estimate the number of participants needed to detect the effects of an intervention. However, power analyses are only as valid as the parameter estimates used. One such parameter, the expected effect size, can vary greatly depending on several study characteristics, including the nature of the intervention, developer of the outcome measure, and age of the participants. Researchers should understand this variation when designing studies. Our meta-analysis examines the relationship between science education intervention effect sizes and a host of study characteristics, allowing primary researchers to access better estimates of effect sizes for a priori power analyses. The results of this meta-analysis also support programmatic decisions by setting realistic expectations about the typical magnitude of impacts for science education interventions.*

Keywords: *effect size, meta-analysis, program evaluation, science education, statistics, student achievement*

In the past two decades, there has been extensive focus on how to calculate power for cluster-randomized trials, or CRTs (e.g., Bloom, 2005; Bloom, Bos, & Lee, 1999; Donner & Klar, 2000; Hedges & Rhoads, 2009; Konstantopolous, 2008; Murray, 1998; Raudenbush, 1997; Raudenbush & Liu, 2000; Raudenbush, Martinez, & Spybrook, 2007; Schochet, 2008). Adequate statistical power for intervention studies helps researchers avoid making Type II errors—errors in which a researcher fails to detect an effect in a sample where that effect indeed exists in the population. Type II errors can occur when there is an insufficient number of participants in the study and/or the effect is smaller than expected. Underpowered studies can lead to inconclusive results that inhibit knowledge accumulation in a field, particularly when the same inconclusive findings are cited repeatedly.

Intervention research in science education is in its infancy in comparison to other fields such as mathematics and reading. Small studies and studies lacking comparison groups abound. In this study, we identified fewer than 2% of 6,600 reports since 2001 that reported impacts from a design conducive to generating confident causal inferences and statistical conclusions (e.g., had a comparison group, included at least 60 participants). From this result, it follows naturally that replications of rigorous impact studies are also rare (Makel & Plucker, 2014; Taylor et al., 2016).

While it is the case that meta-analyses can detect the existence of a significant overall effect from a set of nonsignificant impact estimates, this is contingent on whether there are sufficient studies of an intervention for this meta-analytic utility to come to bear. A lack of conclusive and unbiased findings from primary studies can severely inhibit knowledge accumulation in science education. With so few studies of intervention impacts, the science education field needs more than replications and meta-analyses. Science education needs more sufficiently powered primary studies of program impacts. Once strong causal impact studies with promising effects emerge, replications and meta-analyses will be able to advance the science education field even further.

In this paper, we present the results of a meta-analysis designed to support a priori power analyses in science education research as well as policy or programmatic decisions about intervention effects more broadly. Primary researchers can use the effect size estimates generated by our work in combination with recently published estimates for the intraclass correlation (ICC) and covariate correlation (e.g., Spybrook, Westine, & Taylor, 2016, Westine, Spybrook, & Taylor, 2013) to inform the designs of their own causal impact studies in science education. Policymakers and other decision makers can use our estimates to develop realistic expectations about the types of effects to expect from interventions with specified characteristics. We note here, however, that the most reliable sources of information about the effectiveness of a given intervention are impact estimates from prior studies of that intervention and/or meta-analyses of effects from that or similar interventions. The parameter estimates from this study are meant to refine effect size expectations beyond these primary sources of evidence or provide general guidance in the absence of any extant effect size information.

Although we focus here solely on the effects of science education interventions, this meta-analysis has characteristics that make its findings informative to researchers and decision makers in other education disciplines. This study examines the relationship between the effect size magnitude and key study characteristics, including students' grade level, design of the study, outcome measure type, or intervention focus. Researchers synthesizing intervention studies outside of science education have found noteworthy variation in these very same intervention study characteristics (Cheung & Slavin, 2016; Hill, Bloom, Black, & Lipsey, 2008). As such, we assert that our findings are unique but likely cumulative with those of syntheses outside of science education.

Researchers often use a priori power analyses to estimate the number of participants needed in a study. In cluster randomized trials, accurate a priori power analyses rely on having accurate estimates of the three key design parameters mentioned previously: the ICC (a measure of the between-cluster variance as a fraction of the total variance), the extent to which covariates can account for variation in the outcome, and the estimated effect size. Any a priori power analysis is only as accurate as the design parameter estimates. If any of the design parameter estimates are inaccurate, then too many or too few subjects may be recruited, resulting in higher than necessary costs or an underpowered trial.

Recognizing the importance of accurate power analysis parameter estimates and finding none in the science education literature, the BSCS Science Learning and Western Michigan University began a joint project to provide empirical estimates of these design parameters. Spybrook et al. (2016), Westine et al. (2013), and Westine (2016) are products of this collaboration and examined the ICC and the variance explained by covariates. This manuscript complements that prior work by examining effect sizes from intervention research in science education and does so in a way that extends the scope and approach used in prior synthesis efforts, most notably, the two syntheses of inquiry-based science instruction (Furtak, Seidel, Iverson, & Briggs, 2012; Minner, Levy, & Century, 2010) and the more broadly focused syntheses of elementary science interventions (Slavin, Lake, Hanley, & Thurston, 2014) and secondary science interventions (Cheung, Slavin, Kim, & Lake, 2016). Our study extends the scope of the two syntheses of inquiry instruction by examining a much broader array of science education interventions. In the latter two syntheses, researchers extracted one effect size per study and excluded studies with researcher-developed outcome measures out of a concern about overalignment of outcome to treatment. Our screening and analytic approach contrasts that of Cheung et al. (2016) and Slavin et al. (2014) by modeling the effects of test developer (possible overalignment) instead of using it to exclude studies and extracting multiple effect sizes per study while accounting for the dependency among those effects.

The primary research question of this meta-analysis is: What is the relationship between the magnitude of the intervention effects and key study characteristics? The study characteristics of interest included the design (randomized studies compared to matched quasi-experimental studies), whether the outcome measure was developed by the study authors, who receives the intervention (e.g., students only, teachers only, both students and teachers), the science discipline targeted by the intervention, the treatment provider's role (e.g., researcher or teacher), and the grade level of the students. Together, the suite of papers provides essential information for study designers in science education to conduct a priori power analyses.

## Method

### *Eligibility Criteria*

Eligible interventions were those implemented in either formal education settings or education lab research settings with students in primary and secondary schools. We included only studies published in English because we lacked capacity for translation. We included an array of science achievement outcome measures, including content knowledge, use of scientific practices, and outcomes related to understanding of the nature of science. Eligible interventions are school-based or lab-based interventions of any duration whose efficacy was reported between 2001 and 2014. We selected 2001 as the start of our collection of studies because it was the year of the passage of the No Child Left Behind Act—an act that called attention to the need for experimental intervention research in education and established funding mechanisms for conducting such studies.

We define lab-based interventions as those delivered to students at a university or other research site (e.g., nonprofit site). Interventions in museums only met eligibility requirements if the instruction was formalized. For example, a study of formal lecture demonstrations to classes of students at a museum was included, but studies of free-choice learning at a museum were not included. We included no studies from homeschool settings.

Eligible interventions included curriculum programs (including computer software activities for students to use), professional development programs with student-level outcomes, and use of specific instructional approaches or teaching strategies. Eligible comparison conditions include actual control groups (no intervention), "business-as-usual" (BaU) comparison groups (extant programs or practices), or a sham treatment (watching a movie that was not expected to be particularly beneficial). We did not include alternative interventions as eligible comparison groups. Our decision to exclude treatment-treatment studies was based on the fact that effect sizes from these studies would likely be smaller than (incomparable to) treatment-control, treatment-BaU, or treatment–sham treatment studies. To be eligible, the study design had to include at least two groups whose outcomes could be compared. We included studies in which group assignment was determined randomly (person-randomized or cluster randomized) or nonrandomly (e.g., quasi-experiment). Quasi-experiments were only included when they had clear matching on pretests prior to assignment to treatment or comparison conditions. To limit risk of bias, we required studies to have at least 30 students assigned to either the treatment or comparison group. This decision was informed by the work of Turner, Bird, and Higgins (2013), who concluded that in meta-analysis with at least two well-powered studies (like this one), underpowered studies contribute little additional insight. This finding is further supported by more recent work associating small studies with increased heterogeneity (IntHout, Ioannidis, Borm, & Goeman, 2015) and risk of bias (Afshari & Wetterslev, 2015).

In summary, each study in the ultimate set of eligible studies met all the following criteria:

- based either in a school or an educational research lab setting
- included either primary or secondary students
- published in English
- included at least one student achievement outcome
- published between 2001 and 2014
- studied a specific science education intervention
- included at least two groups that could be compared on the outcomes (i.e., treatment-control, treatment–business as usual, treatment–sham treatment)
- included a pretest or other measure to estimate baseline equivalence
- included random assignment or matching on pretest for quasi-experiments
- included at least 30 students in each treatment or comparison group
- focused on a general population of students (e.g., did not have an explicit focus on students with learning disabilities).

### *Information Sources and Search Strategies*

The study used several information sources to locate qualifying studies: Ulrich's Web (ulrichsweb.serialssolutions.com), the Web of Science database (www.isiknowledge.com), the *ProQuest Dissertation Abstracts* database (www.proquest.com/products-services/dissertations), and reference lists of eligible studies (i.e., reference harvesting). We also identified organizations that were likely to conduct science education research studies and contacted them about unpublished manuscripts (see Search Strategies for Unpublished Studies, in the following).

*Search strategies for published studies.* We used Ulrich's Web (the online version of *Ulrich's Periodicals Directory*) to search for education journals published in English that focused on science education and education research. Specifically, we used the subjects: *Education–Teaching Methods and Curriculum and Sciences: Comprehensive Works* and English language as a delimiter to search Ulrich's Web. The resulting list included 23 journals that publish science education research (see list in online Supplemental Appendix S1).

To search within these journals, we created a search string for use in Web of Science that we intended to capture a wide range of experimental and quasi-experimental studies in science education. The search string was: [TS= (intervention OR control OR treat* OR Experiment* OR Quasi* OR Effect* OR Compar* OR Trial OR Efficacy OR Random OR Assign*) AND PY = (2001 OR … 2014) AND SO= ("Journal of Research in Science Teaching" OR…)] with the additional 22 journals not shown here. We chose to do a journal-specific search because our search was so broad—our original open searches were leading to hundreds of thousands of abstracts. We followed our initial journal-specific search for published studies by examining references of included studies. This reference harvest led to studies from a total of 71 journals (see journal list in online Supplemental Appendix S2).

*Search strategy for unpublished studies (grey literature).* Our grey literature search included: a search string employed within *ProQuest Dissertation Abstracts*, a request sent to listservs, and a search of the websites of 55 research organizations (see online Supplemental Appendix S4) informed by a list used for a similar purpose in the Science Review Protocol of the What Works Clearinghouse (Institute for Education Sciences, 2012). The full dissertation search string including delimiters is provided in online Supplemental Appendix S3. This search string returned 3,516 dissertations completed between 2001 and 2014 that were potentially eligible. After two screeners independently reviewed the abstracts of these studies, we identified 496 dissertations that met abstract screening criteria, and this subset was included in the full-text screening stage of the study. Combined, the request to listservs and the various research organizations resulted in 30 manuscripts submitted to the research team for full-text screening.

### *Data Collection and Coding*

Researchers coding the full text of manuscripts used a FileMaker Pro database. For each study coded, we obtained a portable document format (PDF) file and embedded a link to the file directly into the database for coding and archival purposes. Coders could highlight the PDFs and make comments. Training involved collaborative coding tasks to establish coding norms and independent coding tasks to estimate intercoder reliability. Across all codes, the average level percentage agreement for independent coding was 83%. The coding team held weekly meetings to ask questions and resolve issues that had arisen during the previous week. When discrepancies arose, the PIs made final coding decisions. The database was hosted on a server such that all team members could access the database simultaneously.

Key information about the characteristics of a study or statistical information needed to extract effect sizes was often missing from published reports. In these situations, coders contacted study authors directly to inquire about the missing information. Specifically, 59 study authors were queried requesting various information, including intervention dosage, demographic information, details about the instrument used in study outcomes, timing of the posttest, the type of assignment to groups (e.g., RCT), and requests for means and standard deviations used in analyses. Most often, requests were for more specific information about the frequency or duration of the intervention (dosage), requested in 24% of the queries, and demographic information disaggregated by treatment group, requested in 20% of the queries. Nineteen percent of queried authors responded, sharing important information relevant to coding the studies. In the remaining instances, authors no longer had access to the data or did not respond to our query.

*Variables coded.* In addition to the bibliography and eligibility tables, there were four tables in the database: header (data related to the study as a whole), groups (data about each treatment and comparison group in the study), dependent variables (data about each outcome variable in the study), and effect sizes (data about group means, standard deviations, observed sample sizes, and/or other information used to estimate effect sizes, including *t* tests, *p* values, or hand-calculated effect sizes). Each table included a space for notes related to coding problems. Studies were required to pass criteria from the full-text eligibility table first before additional coding ensued. These initial eligibility requirements entailed coder assessments of whether the intervention was in science education and in either a school- or lab-based setting and the study occurred in a relevant timeframe (after 2001) with a relevant outcome and was conducted with a policy-relevant K–12 student sample and met methodological requirements, such as using an eligible comparison group and meeting a minimum sample size threshold, and reported impacts in such a way that an effect size could be extracted. See specifics for these codes in online Supplemental Table S1.

*Choice and coding of moderators.* Our choices of effect size moderators to test were influenced by our desire to provide evidence that either corroborates or challenges findings in the extant literature. For example, Hill et al. (2008) examined the extent to which the nature of the outcome measure (broadly focused standardized test vs. specialized topical test) and the grade level of the students (elementary school vs. middle school vs. high school) influenced effect size magnitudes from multiple disciplines, finding larger effect sizes for specialized topical tests and middle school interventions. Extending these analyses using a multiple regression approach (meta-regression), Cheung and Slavin (2016) also tested the effects of outcome measure type and grade level, finding larger effect sizes in studies with researcher-made measures and studies of elementary school students. Additionally, Cheung and Slavin (2016) tested the effects of study design (RCT vs. QED). The approach in the current study sought to build on these prior efforts.

We primarily used binary codes but in one case used a set of related indicator (dummy) codes. *RCT* is a binary study design variable designating randomized (coded 1) or quasi-experimental studies (coded 0). *TCHONLY* and *TCH&STU* are dummy variables that indicate who received the intervention. In this coding scheme, the reference group (*STUONLY)* includes interventions received only by students (e.g., curriculum materials without professional development support for teachers). *TCHONLY* indicates that the intervention is delivered only to teachers (e.g., a professional development program; coded 1 if it is teacher only and 0 otherwise), and *TCH&STU* indicates that the intervention is delivered to both teachers and students (e.g., curriculum materials plus supporting professional development for teachers; coded 1 if it is an intervention for both teachers and students and 0 otherwise). The *SCITYPE* variable is a binary variable that describes the science discipline under investigation. We grouped life science (including biology), multidisciplinary science, and earth/space science together (coded

0) as science disciplines that generally do not include mathematics at the high school level and grouped physical science, physics, and chemistry together (coded 1) as science disciplines that do generally involve mathematics. *TESTDEV* reflects whether an outcome was developed by study authors (coded 1 for author developed assessments, 0 otherwise). The *HIGRD* variable is a binary variable with 0 representing primary and middle school students (K–8 in the U.S. context) and 1 representing upper secondary school students (9–12 in the United States). Finally, the *TRTPROV* variable is a binary variable with 1 indicating that the teacher was the intervention provider and 0 indicating that the intervention provider was someone other than the teacher (e.g., a researcher).

*Inclusion of multiple effect sizes.* In traditional meta-analyses, each study contributes one effect size and the effect size estimates are independent across studies. However, researchers frequently report multiple effects. In some instances, researchers might use multiple outcome measures with the same individuals (e.g., one assessment might measure students' understanding of science content, and another might measure students' understanding of scientific practices such as the development of explanations). In other instances, multiple outcome measures arise from testing the same students at multiple timepoints (posttest and delayed posttest models fall in this category). Some studies have two or three treatment groups and each treatment group is compared to the same comparison group.

*Effect size calculation.* The 636 effect sizes (Hedges' *g*) extracted from the eligible studies were a combination of pretest and posttest effect sizes. Of the 636 effects, 292 were posttest effect sizes that were the focus of our analysis. The magnitude and variance of each of the 292 effect sizes are provided in online Supplemental Table S2. Of the 292 effect sizes, 220 were calculated using *unadjusted* means but had a pretest effect size for the same dependent variable and treatment/comparison group contrast. Another 72 effect sizes were calculated using means adjusted for covariates. For the 220 studies with a matched set of pretest and posttest effect sizes for each outcome, we calculated the difference between effect sizes (posttest treatment effect minus pretest treatment effect). All effect sizes, including those from cluster-randomized or cluster quasi-experimental designs, were standardized on the individual-level, treatment group–specific, standard deviations and sample sizes.

Frequently, studies using cluster-randomized or cluster quasi-experimental designs failed to use appropriate analyses. That is, the study used cluster assignment (e.g., entire classes or schools were assigned to treatment or comparison conditions), but the analyses did not account for clustering (analyses were conducted as if individual students were assigned to treatment or comparison conditions). In such circumstances, the reported sample size is too large, and the analyses use underestimated standard errors. Higgins and Green (2011) describe how meta-analysts can adjust for such mismatched analyses by calculating a reduced, effective sample size for the study. Equation 1 conducts this adjustment:

$$N_{adj} = N / [1 + (M - 1)ICC], \tag{1}$$

where *M* is the average cluster size for the study, *N* is the original number of student participants, and ICC is the intraclass correlation. We used a value of .172 for the ICC as this value is the eighth-grade science ICC estimate from Westine et al. (2013) and was chosen because it approximates a middle school ICC and thus a median position along the K–12 continuum.

If the number of clusters for a given group was 1 (e.g., cluster assignment with one treatment class and one comparison class), no adjustment is made. This design feature occurred in 15 studies, and we report results of sensitivity analyses in online Table S3 that compare results with these studies included versus omitted. In addition, we did not adjust the number of participants for studies in which individual students were assigned to the treatment or comparison condition.

After we adjusted the number of participants based on cluster size, we Winsorized the effective sample sizes (N) computed in equation 1. We used the Winsorized values of *N* along with the standardized mean difference effect sizes to calculate Hedges' *g* effect sizes for each study. Finally, we Winsorized the Hedges' *g* effect sizes. We used Winsorized values so particularly large studies and studies with particularly large effect sizes did not have a disproportionate influence on the results.

*Statistical model.* Our data included 636 effect sizes across 96 studies (an average of 6.6 effect sizes per study). Effect sizes within studies are correlated, and this dependence needs to be accounted for in the analysis. As information on the correlation between these effect sizes was not reported in primary studies, we used a method that was robust to misspecification of the correlation structure. For this reason, we used weighted least squares to estimate the meta-regression model and adjusted the standard errors for dependence within studies through use of robust variance estimation (RVE; Hedges, Tipton, & Johnson, 2010). Unlike standard model-based methods such as multivariate meta-analysis (Jackson, Riley, & White, 2011), RVE does not require the correlation structure to be correctly specified when calculating standard errors and hypothesis tests; instead, it estimates the standard errors empirically using a sandwich-estimator (for a tutorial, see Tanner-Smith, Tipton, & Polanin, 2016).

Additionally, we use small-sample corrections to RVE (Tipton, 2015; Tipton & Pustejovsky, 2015). These small-sample corrections involve the specification of a "working model" for the correlation structure—here we use a "correlated effects" model—and use of a *t* distribution with Satterthwaite degrees for the reference distribution. These degrees of freedom are a function of both the number of studies and features of the covariates. Importantly, these small-sample adjustments allow RVE to appropriately account for dependence even when degrees of freedom are as small as 4 or 5. Even though a working model is specified, a wide range of simulations shows that the resulting standard errors and hypothesis tests are robust to misspecification (Tipton, 2015; Tipton & Pustejovsky, 2015).

*Working model and weighting.* In most studies in this meta-analysis, the effect sizes were dependent because they were measured on the same individuals. For this reason, we assumed the "correlated effects" model as our working model in RVE. We also used this model to define approximately inverse-variance weights, defined as

$$w_{ij} = \frac{1}{k_j(\bar{v}_{\bullet j} + \hat{\tau}^2)}, \tag{2}$$

where $w_{ij}$ is the weight for effect size *i* in study *j*, $\hat{\tau}^2$ is the between-study random effect estimated using Equation 15 from Hedges et al. (2010), $\bar{v}_{\bullet j}$ is the unweighted average of the variances of the effect size estimates in the *j*th study, and $k_j$ is the number of effect sizes from study *j*.

*Meta-regression.* Our meta-regression model for the primary analysis was estimated using the R package *robumeta* (Fisher, Tipton, & Hou, 2017) and specified as:

$$\begin{aligned} ES_{ij} = {} & \beta_0 + \beta_1(RCT)_{ij} + \beta_2(TCHONLY)_{ij} \\ & + \beta_3(TCH\ \&\ STU)_{ij} + \beta_4(SCITYPE)_{ij} \\ & + \beta_5(TESTDEV)_{ij} + \beta_6(HIGRD)_{ij} \\ & + \beta_7(TRTPROV)_{ij} + e_{ij}. \end{aligned} \tag{3}$$

We grand mean centered all variables in the regression. As a result, the intercept is an estimate of the average effect size at the grand mean of all predictors. We assessed statistical significance of an estimated regression coefficient ($\beta_j$) with the test statistic $t_j^R$:

$$t_j^R = \frac{b_j}{S_j^R}, \tag{4}$$

where $S_j^R$ is the robust standard error including adjustments for small samples (Tipton, 2015).

The test statistic is compared with critical values from student's *t* distribution with $\eta_j$ degrees of freedom (these degrees of freedom are estimated and can vary from covariate to covariate; see Tipton, 2015).

*Missing data handling* We were unable to determine the treatment provider's discipline (teacher or other) for 16 effect sizes across four studies. Rather than lose the data completely, we conducted a single imputation and ran the meta-regression on a complete data set. We used *MPlus* to run a multilevel imputation to account for the nested structure of the data. We conducted a sensitivity analysis, comparing the values of the coefficients for the unimputed data set to those from the imputed data set and found little difference. The largest difference was in the coefficient for *TRTPROV*, with $\hat{\beta}$ = .040 for the unimputed set and $\hat{\beta}$ = .013 for the imputed set (with virtually identical standard errors). Taking this one step further, we tested whether the results of our single imputation were stable over 10 imputations and found very little difference in estimates across imputations (see online Supplemental Table S3 for details).

## Results

### Study Selection

The abstract screen culled the list to 1,174 unique studies that underwent a full-text screen. The full-text screen reduced the number of eligible studies to 96, and these studies came from a total of 21 different countries: United States (43), Turkey (15), Israel (7), Taiwan (6), Canada (3), Germany (3), Nigeria (3), Jordan (2), Kenya (2), Brazil (1), China (1), England (1), Finland (1), Greece (1), India (1), Korea (1), Netherlands (1), New Zealand (1), Singapore (1), Slovenia (1), and Switzerland (1). Figure 1 identifies the number of studies at each stage of our process. In this figure, a "record" refers to a single citation from one database. Because we used multiple databases, sometimes different databases located the same record. We identified 6,622 records through database searching (raw number, including duplicates), along with 30 additional records through our grey literature search. The combined total number of unique records was just 6,637 after excluding 15 duplicates. We use the term *articles* to refer to unique records. Our abstract screen of 6,637 articles led to the inclusion of 1,286 articles that met our abstract screen criteria.

It was essential that coders not include effect sizes for the same outcomes from the same participants more than once. This problem can arise when authors publish more than one article from the same research study. We searched author names across all 1,286 articles to look for articles that appeared to report findings from the same participants. We combined multiple articles into a single "study" and coded at the study level. The process of linking related articles into unique studies brought our number of studies down to 1,174.

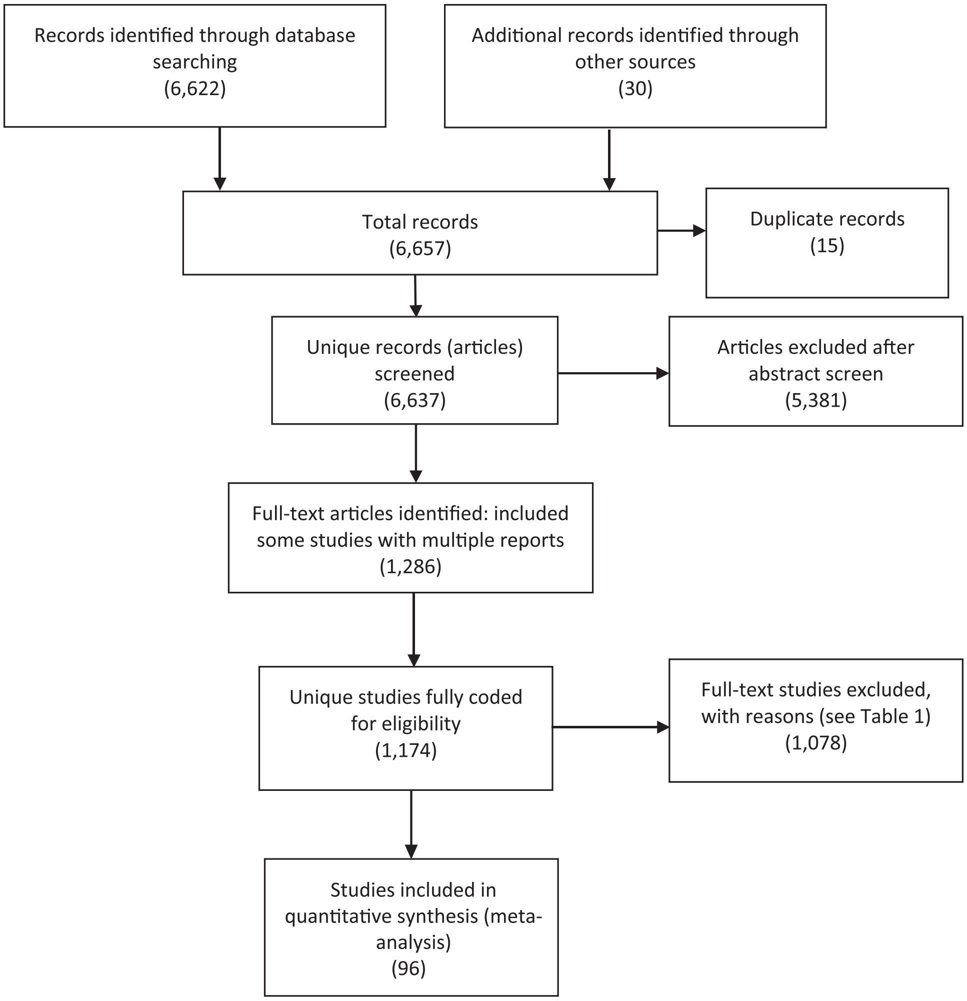
Records identified through database searching
(6,622)
Additional records identified through other sources
(30)
Total records
(6,657)
Duplicate records
(15)
Unique records (articles) screened
(6,637)
Articles excluded after abstract screen
(5,381)
Full-text articles identified: included some studies with multiple reports
(1,286)
Unique studies fully coded for eligibility
(1,174)
Full-text studies excluded, with reasons (see Table 1)
(1,078)
Studies included in quantitative synthesis (meta-analysis)
(96)

FIGURE 1. *Number of records, articles, and studies identified and retained at each stage of the selection process.*

Of these, a full-text eligibility screen (using information from all linked articles for each study) yielded 96 studies that met our full inclusion criteria.

We excluded studies during the full-text screen for a variety of reasons. Many abstracts were sufficiently vague that we included the associated articles in the full-text screen for no reason other than there was insufficient information in the abstract to make an eligibility determination. For example, an abstract might refer to "impacts of an intervention" but not mention the existence or absence of a comparison group. Other abstracts might refer to "students" without clarifying (in the title, abstract, or keywords) whether they were primary, secondary, or undergraduate students.

Occasionally, study coders continued coding beyond identifying a reason for excluding a study. This occurred particularly when coders were unsure of a disqualifying coding decision and sought consultation with the larger team before excluding a study. Thus, some studies had multiple reasons listed for their exclusion while others had just one. Table 1 provides a summary of the reasons studies were excluded. However, without further coding of all disqualifying features of every study, the percentages should not be interpreted as complete; in fact, they likely underestimate the number of studies lacking a characteristic.

As the table shows, the two most common reasons for excluding a study related to study design. Twenty-two percent of the excluded studies did not include an eligible comparison group (this includes studies that had no comparison group as well as those that had an alternative treatment comparison group). We excluded 21% of studies because the student sample size was too small (fewer than 30 students in either the treatment or comparison group). A substantial portion (14.7%) were excluded because there were no reported measures of baseline equivalence. We included studies provided that the baseline measures were reported but did not use magnitude of the baseline equivalence effect size as an exclusion criterion.

TABLE 1

*Reasons Studies Were Excluded From Meta-Analysis*

| Reason for exclusion | Number excluded | Percentage excluded |
|---|---|---|
| Did not include a science education intervention | 147 | 13.6 |
| Neither school-based nor education research lab–based | 25 | 2.3 |
| Did not include an eligible student outcome | 74 | 6.9 |
| Did not include at least one eligible comparison group | 237 | 22.0 |
| Not conducted with primary or secondary students | 125 | 11.6 |
| Not published in or after 2001 | 6 | 0.6 |
| Did not have at least 30 students per treatment group | 228 | 21.2 |
| Did not include a measure of baseline equivalence | 158 | 14.7 |
| Did not include any achievement outcome | 34 | 3.2 |
| Included only special populations | 23 | 2.1 |
| A quasi-experiment without pretest matching | 110 | 10.2 |
| Did not provide sufficient information for coding effect sizes | 10 | 0.9 |

TABLE 2

*Mean Effect Size by Study Design or Intervention Characteristic*

| Variable | Reference group (*n* = No. of ES) | Mean ES (*SD*) Reference group | No. of studies in reference group | Indicator group (*n* = No. of ES) | Mean ES (*SD*) Indicator group | No. of studies in indicator group |
|---|---|---|---|---|---|---|
| *RCT* | Matched quasi-experiments (*n* = 50) | 0.49 (0.32) | 9 | RCTs (*n* = 242) | 0.45 (0.48) | 87 |
| *STUONLY* |  | NA | NA | Interventions for students only (*n* = 59) | 0.36 (0.46) | 24 |
| *TCHONLY* |  | NA | NA | Interventions for teachers only (*n* = 29) | 0.47 (0.28) | 6 |
| *TCH&STU* |  | NA | NA | Interventions for both teachers and students (*n* = 204) | 0.58 (0.48) | 66 |
| *SCITYPE* | Life, earth/space, and multidisciplinary science (*n* = 155) | 0.49 (0.42) | 49 | Physics, chemistry, and physical science (*n* = 137) | 0.45 (0.46) | 47 |
| *TESTDEV* | Assessment developer is not the author or researcher (*n* = 79) | 0.29 (0.43) | 32 | Assessment developer is the author or researcher (*n* = 213) | 0.67 (0.47) | 64 |
| *HIGRD* | Highest grade in study includes primary and lower secondary students (K–8) (*n* = 102) | 0.41 (0.47) | 41 | Highest grade in study includes upper secondary students (9–12) (*n* = 190) | 0.53 (0.45) | 55 |
| *TRTPROV* | Treatment provider is not a teacher (*n* = 64) | 0.49 (0.44) | 20 | Treatment provider is a teacher (*n* = 212) | 0.45 (0.45) | 76 |

### *Overall Statistics*

Table 2 shows the mean posttest effect size, standard deviation, and study sample size for each category of study design or intervention characteristic. The weighted average pretest effect size was −0.01 standard deviations, suggesting that although we did not disqualify studies based on baseline equivalence, it does not appear to have introduced substantial bias in the impact estimates.

In addition, we report several other statistics of importance. The intercept ($\hat{\beta}_0$ = 0.489) of the meta-regression provides an estimate of the overall mean effect size for the study sample, and the 95% confidence interval around this intercept [0.368, 0.610] provides a sense of the precision of this estimate.

The overall average effect, however, is not particularly useful when there is heterogeneity in effect sizes across studies.

TABLE 3
*Results of Robust Variance Estimation Meta-Regression Using 96 Studies With 292 Effect Sizes*

| Parameter | Estimate | *SE* | *t* value | *df* | *p* | 95% CI Lower | 95% CI Upper |
|---|---|---|---|---|---|---|---|
| Intercept | 0.489 | 0.060 | 8.161 | 43 | <.001 | 0.368 | 0.610 |
| *RCT* | −0.004 | 0.179 | −0.024 | 11 | .98 | −0.397 | 0.388 |
| *TCHONLY* | 0.060 | 0.148 | 0.405 | 9 | .69 | −0.276 | 0.396 |
| *TCH&STU* | 0.149 | 0.136 | 1.097 | 40 | .28 | −0.125 | 0.423 |
| *SCITYPE* | −0.044 | 0.120 | −0.363 | 61 | .72 | −0.284 | 0.196 |
| *TESTDEV* | 0.258 | 0.102 | 2.522 | 48 | .02 | 0.052 | 0.463 |
| *HIGRD* | 0.053 | 0.114 | 0.463 | 68 | .65 | −0.175 | 0.280 |
| *TRTPROV* | −0.005 | 0.139 | −0.036 | 22 | .97 | −0.295 | 0.284 |

*Note. RCT*, 1 = randomized design; *TCHONLY*, 1 = intervention with teachers only; *TCH&STU*, 1 = intervention for both teachers and students; *SCITYPE*, 1 = physics, chemistry, or physical science; *TESTDEV*, 1 = assessment developer is author or researcher; *HIGRD*, 1 = highest grade in study includes upper secondary (Grades 9–12); *TRTPROV*, 1 = treatment provider is a teacher. See Table 2 for the reference group of each variable.

To understand the true variation in study-average effect sizes, a prediction interval is useful. In this study, the 95% prediction interval for study-average effect sizes is [–0.393, 1.371], indicating that while the average study produced a positive effect, in some studies, the true effect was negative (and in others, the effect was positive and much larger). This interval is based on the fact that the variation in effect sizes across studies was high ($I^2 = 77.36\%$, $\tau = 0.45$).

### *Study and Intervention Characteristics as Moderators*

*Moderator effects from the meta-regression.* Table 3 shows the parameter estimates from our meta-regression using RVE with correlated effects weights, conducted using Equation 3. Note that sensitivity analyses for the effect of Winsorizing (vs. not Winsorizing) the effect sizes indicated no difference in any parameter estimate to three decimal places. Sensitivity analyses were also conducted with regard to including studies with a single cluster per treatment condition. Two noteworthy differences emerged, and these results are reported and discussed in Table S3 online.

The interpretation of the meta-regression coefficients (slopes) is conceptually similar to any multiple regression using binary indicators—each regression coefficient is a covariate-adjusted difference in mean effect size between groups of effect sizes that differ on the target characteristic. For example, the regression coefficient of −0.004 for RCT represents the model-based estimate of the difference in average effect size for RCTs compared to QEDs. That is, controlling for all covariates, RCT effect sizes are estimated to be 0.004 standard deviations smaller than that of matched QEDs, on average. Similarly, the effects from researcher-developed assessments were estimated to be nearly 0.258 *SD* larger, on average, controlling for other study, sample, and outcome characteristics. This effect was statistically significant.

Also notable (though not significant) is that interventions with components for both students and teachers produced higher effect sizes than those that target students only with an adjusted difference of 0.149 standard deviations. When an intervention was conducted in a science discipline that tends to be more mathematical, the adjusted effect sizes were slightly smaller (0.044 standard deviations) than those in science disciplines that are less mathematical. Interventions for secondary students (Grades 9–12) showed slightly higher effects than for students in primary and lower secondary (K–8) grades, with an adjusted difference of 0.053 standard deviations. Finally, when teachers delivered an intervention, the effects were nearly identical (adjusted difference = 0.005 *SD*) to those from interventions delivered by a researcher (or other nonteaching personnel). Note that we also explored publication bias using a similar analytic approach, and these ancillary results are provided in online Supplemental Appendix S6.

### *Using the Meta-Regression Parameter Estimates in A Priori Power Analyses*

When conducting a priori power analyses, a study designer should first consider whether there are existing meta-analyses or effect sizes from isolated primary studies of the same or similar interventions. In the absence of such information, we propose use of our meta-regression parameter estimates to arrive at an empirically based effect size estimate. The least precise approach would be to use the intercept estimate, which provides the overall average (weighted) of all 292 effect sizes in the meta-analysis. We don’t anticipate that this would be appropriate in the majority of cases as most study designers will have information on at least a subset of the study and/or outcome characteristics that can be used to adjust the overall mean effect size estimate. Optimally, a study designer would have information on all variables and make the eight corresponding adjustments to the grand mean, based on the magnitude of the meta-regression coefficients in Table 3. Yet another variation that leverages existing impact information would be to use an extant impact estimate or summary effect from a meta-analysis in place of the intercept from our model and then make the corresponding adjustments using our meta-regression coefficients.

To facilitate accurate and convenient computation of predicted effect sizes based on selected study/outcome characteristics, we developed a Web-based application that uses the results generated by the R *robumeta* and *clubSandwich* (Pustejovsky, 2015) packages. The application can be found at https://effectsizecalculator.bscs.org. For details on how the online application computes predicted effect sizes from a combination of the parameter estimates in Table 3 and user input, see the online Supplemental Appendix S5.

*Example: Using meta-regression estimates to estimate an effect size.* Here we present an example of how science education researchers may use our meta-regression results in a power analysis. Assume a researcher has developed a high school (*HIGRD* = 1) biology intervention (*SCITYPE* = 0) that provides curriculum materials for students and professional development for teachers (*TCHONLY* = 0; *TCH&STU* = 1), and the teachers implement the curriculum materials with students (*TRTPROV* = 1). The researcher wants to conduct a cluster (i.e., school) randomized efficacy trial (*RCT* = 1) and would like to determine how many schools are needed for the study to achieve 80% power. The researcher plans to use a state standardized test (*TESTDEV* = 0) as an outcome measure for the efficacy trial. Entering these study and outcome characteristics into the online application yields a predicted effect size of 0.377.

In addition to reporting an estimated effect size, the online application also indicates the precision of this estimate. A 95% confidence interval is computed for each expected effect size based on the set of observed covariate values. These intervals are calculated using

$$\text{Est} +/- \text{t}\left(.025, \eta\right)\text{SE}\left(\text{Est}\right), \qquad (5)$$

where both the degrees of freedom ($\eta$) and standard error (i.e., *SE*[Est]) are estimated based on the small-sample *F* test (Tipton & Pustejovsky, 2015). Importantly, the degrees of freedom $\eta$ here will differ for different predicted effect sizes. For the aforementioned example, the application computes a 95% confidence interval around the predicted effect size (0.377) with a lower and upper limit of 0.164 and 0.590 standard deviations, respectively.

Using findings from Westine et al. (2013), study planners can obtain a reasonable estimate of the school-level intraclass correlation for 10th-grade outcomes (ICC = 0.196) and the variance explained by a school-level pretest covariate ($R^2$ = 0.868). Combining this information with the predicted effect size of 0.377 standard deviations and using Optimal Design v. 3.1 (Spybrook et al., 2011), one finds that a cluster randomized trial must maintain an analytic sample of 13 schools with 50 students each to achieve 80% power (5% significance level). Similarly, although the merits of doing so are debatable, a study designer could choose to take either an ultraconservative or ultraliberal approach and use the lower (conservative) or upper (liberal) limit of the effect size confidence interval in the power analysis. Using the same values as previously described for the $R^2$, ICC, cluster size, significance level, and desired power, Optimal Design yields an analytic sample estimate of 52 schools using the lower limit and 8 schools using the upper limit. Given this wide range, we would recommend instead using extant effect size information (if available) from prior primary studies or meta-analyses of the intervention or similar interventions.

## Discussion

### *Comparisons to Extant Research*

*The average effect size (the intercept).* The grand mean effect size we estimated for this study, 0.489, is larger than any of the sample size weighted mean effect sizes reported in recent synthesis studies of science education interventions. Two recent studies are particularly relevant to the present study. The first, Slavin et al. (2014), synthesized a total of 23 effect sizes for elementary school science interventions, finding the following summary effects for three intervention categories: 0.02 *SD* for science kits, 0.36 *SD* for professional development programs, and 0.42 *SD* for technology applications. Similarly, Cheung et al. (2016) synthesized 21 effects from secondary school science interventions, finding the following average effects across four intervention categories: 0.17 *SD* for instructional process programs, 0.47 *SD* for technology programs, –0.02 *SD* for science kits, and 0.10 *SD* for innovative textbooks, each smaller than the grand mean effect size of this study. The differences between our results and that of other recent studies can be attributed in part to differences in the sample of effect sizes synthesized. The studies described previously synthesized a total of 44 effect sizes, as opposed to the 292 in the current study. The sample differences between the prior research and the current study arise from differences in inclusion criteria (e.g., publication date, minimum intervention duration, and the eligibility of studies that used experimenter-developed outcome assessments) as well as differences in analytic approach. Regarding the analytic approach, the studies described previously used fixed effects variances and traditional meta-analytic approaches to estimate summary effect sizes, whereas the present study extracted multiple effect sizes per study and estimated adjusted mean effects using meta-regression that modeled within-study dependence of effect sizes through robust variance estimation.

*The effect of study design.* The effect of study design was very small but consistent with recent work in this area. Although we observed a smaller effect of design than that observed by Cheung and Slavin (2016), the direction of the effects is the same, with both studies estimating smaller effects for randomized designs.

*The effect of bundled interventions.* The results of this analysis are suggestive that there is a positive effect of developing bundled interventions that provide products and/or services for both teachers and students. Unfortunately, this tentative result cannot be corroborated by the two recent syntheses by Slavin and colleagues as they categorized interventions differently than we did in the present study.

*The effect of science discipline type.* The effect of discipline type was quite small, and the possibility that this effect is spurious is too high to support a confident claim. Further study is needed around whether a noteworthy effect of discipline exists in the effect size population.

*The effect of who develops the outcome measure.* The finding that stands out dramatically is the positive relationship between use of researcher-developed outcome assessments and the magnitude of the treatment effect. This can result from either overalignment of outcome measures to treatments, insensitivity of standardized measures to treatment effects, or both. We cautiously assert that the primary source of this relationship in our data is likely the tendency of broadly focused standardized assessments to be insensitive to treatment effects. We found in our coding of each study's methodological approach only a few instances of treatment-outcome overalignment but acknowledge that assessments of overalignment can be subjective and no clear definition exists.

*The effect of students' grade level.* Slavin et al. (2014) did not report an overall summary effect for the 23 effect sizes in their synthesis of elementary school science interventions, nor did Cheung et al. (2016) report the like for the 21 effects of secondary school science interventions in their synthesis. However, both studies reported the individual study effects and sample sizes necessary to compute overall summary effects by grade span (elementary vs. secondary). Using this information, we conducted a random effects meta-analysis with student sample size weighting, finding for elementary school interventions a weighted summary effect of 0.33 *SD* and for secondary school interventions, a weighted summary effect of 0.21 *SD*. This finding is consistent with results from Hill et al. (2008), who found larger average effects in the earlier grades for interventions in mathematics and reading.

It would appear that the findings of the present study, where the weighted average of effect sizes for secondary school interventions is higher than for elementary school interventions, diverge from that of the prior syntheses. However, this appears to be an artifact of how the grade intervals were coded in the present study (K–8, 9–12) as opposed to the other syntheses. For example, the grade intervals in the Slavin et al. (2014) and Cheung et al. (2016) studies were K–5 and 6–12, respectively. When we disaggregate our effect sizes into these new grade intervals, we see a similar effect of grade level, with the weighted average effect size for interventions in the K–5 grade interval estimated at 0.09 standard deviations greater than the weighted average effect size for interventions in the Grades 6–12 interval. This is largely consistent with the mean effect size difference across these grade intervals from the Slavin et al. and Cheung et al. work.

*The effect of who delivers the intervention.* The effect of who delivers the intervention was also quite small and inconclusive. Although further study is needed to assess whether an effect of treatment provider exists in the effect size population, the lack of a clear effect challenges conventional wisdom in education that larger effects will be observed (all else equal) when the intervention developer also delivers the intervention (e.g., proof of concept or efficacy studies) as opposed to when a nondeveloper implements the intervention (e.g., scale-up studies).

## *Limitations*

This study had several notable limitations. First, across the set of eligible studies, there was significant imbalance across the categories of several nominal variables. This required us to dichotomize planned moderators to achieve better balance and statistical power or in some cases eliminate the moderator altogether.

Omitted moderators included participant characteristics such as percentage minority, a contrast for lab- versus school-based intervention, and intervention duration. Percentage minority could not be used due to insufficient reporting in general and by treatment condition in particular. The lab- versus school-based intervention contrast could not be used due to extreme imbalance (4 lab-based, 92 school-based interventions). Outcome type could not be used as a variable because too few effect sizes for affective outcomes existed. The moderator for comparison group type suffered the same fate as only 5 studies used a no intervention comparison group, while 91 used a business-as-usual counterfactual. Finally, intervention duration could not be used because the data were extremely skewed and the resulting meta-regression degrees of freedom were less than 4, the minimum cutoff for trustworthy results suggested in the *robumeta* documentation. This was unfortunate as intervention duration is something that study planners do tend to know prior to study implementation, and it seems reasonable to hypothesize that duration would explain variance in effect size magnitude. We acknowledge that the omission of these planned moderators could have introduced unknown confounds in the interpretation of the remaining coefficients. Further, correlations among the remaining moderators could have affected the results. For example, it is intuitive that the *TCHONLY* and *TCH&STU* variables might be correlated, and indeed, the phi coefficient for this pairing is the largest of all bivariate relationships (0.51). However, overall, the relationships among the moderators are not strong, with the average absolute value of all pairwise bivariate correlations (phi) = .14.

The planned moderator that required a collapse into binary categories was *SCITYPE*. The original coding included six categories of science subject matter foci, but there was poor balance across the categories. Across the 96 studies, the percentage of interventions with each subject matter focus was biology/life science (37.8%), chemistry (19.5%), physics (20.7%), earth/space science (3.7%), multidisciplinary science (11%), and physical science (7.3%). Ultimately, we decided to combine physics, chemistry, and physical science into a category of disciplines that tend to be more quantitative at the K–12 level and contrast that with a reference group of traditionally less quantitative disciplines at the K–12 level: biology/life science, earth/space science, and multidisciplinary science. We acknowledge that in doing so, we sacrificed a more fine-grained set of categories and comparisons for better balance and statistical power.

Some unbalanced moderators remained (particularly *RCT*), and the combination of lingering imbalance and parameter estimates that tended to be small in magnitude resulted in poor statistical power and nonsignificant results. Therefore, the coefficients will not support definitive claims or generalizations about the relative efficacy of one intervention approach over another outside of this sample in the larger population of science education intervention effect sizes (with the exception that effect sizes from outcomes developed by the researchers tend to be larger). However, the coefficients provide a useful starting place for a priori power analyses as they provide some empirical basis for design decisions, reflecting variation of effect sizes within this sample of studies. The estimates from this study can also help policymakers and other decision makers in science education begin to set clearer expectations for the magnitude of effects that are likely to be observed from interventions with various characteristics. Finally, this meta-analysis was limited to school- or lab-based interventions. Online interventions and interventions based in informal settings were excluded, limiting the extent to which the study findings apply in an emerging digital age.

## Implications

### *For Power Analyses*

The field of intervention research in science education is in its infancy. After searching over 6,600 abstracts, we found just 96 that met our full inclusion criteria. The lack of a wide research base means that few intervention researchers have data from a comparison group study on which to base their own power analyses for future work. Even when they do, those results may be based on design characteristics that differ in important ways from what a researcher might propose in a larger efficacy trial. Specifically, as Lipsey and Wilson (2001) have suggested, pre-post (one group) effect sizes are incongruent with comparative (two group) effect sizes, although a conversion can be made between the two if the pre-post correlation is known (Borenstein, 2009). As we have shown, overestimating an effect size can lead to drastic underestimation of required sample sizes. When designing studies, a researcher can (and should) use relevant pilot effect sizes, provided that they map to the design of their planned study. Alternatively, a researcher would be wise to use meta-analytic findings (should they exist) for interventions that are a close match to their own invention. Barring that, the use of data from our set of studies would allow researchers who might otherwise have little information on which to base effect sizes for a priori power analyses to make empirically based decisions in this important stage of study design. At a minimum, researchers will have some idea of whether their proposed effect size might be conservative or optimistic based on what we have seen in the science education literature. Using results from our meta-analysis, intervention researchers in science education will be better able to design studies of causal impacts. Better designed studies are more likely to be funded and published. We propose that an important implication of our work will be an improvement in the quality and quantity of impact studies in science education.

Further, the use of adequately powered intervention studies in science education has importance for the field (beyond what it means for individual researchers). The tendency of the field to discount studies with nonsignificant *p* values (and ignore effect sizes), justifiably or not, means that we are likely to overlook important interventions simply because impact studies of those interventions were underpowered. Inconclusive findings can only lead to stagnation of the accumulation of knowledge about science education interventions.

### *For Programmatic Decisions*

Over and above its utility for study designers (i.e., power analyses), the results of this study help establish an effect size "landscape" for science education. The prediction interval described previously establishes a range of plausible values for individual effect sizes, indicating that for interventions like those in this study, 95% of the effect sizes are likely to fall between −0.393 and 1.371 standard deviations. In terms of central tendency, merely interpreting the intercept estimate ($\hat{\beta}_0 = 0.489$) suggests that the average effect size for science education interventions like those in this meta-analysis is about half of a standard deviation. Further, the confidence interval of this intercept [0.368, 0.610] establishes a range of likely values for the overall effect size in the larger effect size population, with the lower limit suggesting the most conservative expectation and the upper limit the most liberal. Beyond these overall findings, relationships observed in the moderator analyses can be helpful to decision makers in education. For example, our largest moderator effects suggest that (all else equal) effects on scores from researcher-developed tests are more likely to be larger than effects on other outcomes measures (e.g., state standardized tests) and that interventions that target both teachers and students (e.g., teacher professional development bundled with curriculum materials for students) are more likely to yield larger effects than interventions that target students only.

### *For Synthesis Methods*

To date, most of the oft-cited work around benchmarks for effect sizes estimates have used a univariate approach to aggregating and summarizing average effect sizes (e.g., Cheung et al., 2016; Hill et al., 2008; Slavin et al., 2014). Considering our findings, we join recent calls to decrease the use of this approach (Polanin & Pigott, 2015). Although in general the adjusted mean differences from our meta-regression coefficients mirror the raw differences in mean effect sizes from the descriptive statistics table (Table 2), both in magnitude and direction, important differences remain. For example, we highlight the *TRTPROV* variable where the difference in raw mean effect size between the set of effects where the teacher provided the intervention are 0.15 standard deviations larger than for the set of effects where a non-teacher provided the intervention. The meta-regression coefficient also represents a mean difference in favor of teacher-provided interventions, but that estimated difference is less than 0.01 standard deviations. Such a difference, when brought into an a priori power analysis, could have a significant influence on minimum sample size estimates or minimum detectable effect sizes. Given the growing literature base suggesting that meta-regression estimates obtained using RVE are trustworthy even with small sample sizes (see simulations in: Tipton, 2013, 2015; Tipton & Pustejovsky, 2015), we conclude that the adjusted mean differences from the meta-regression are more precise and point a way forward for future research in this area.

### *For Future Research*

In this age of evolving technology, science education interventions are becoming more transportable to other formats, contexts, and learning environments. As such, parallel work is desperately needed for online interventions and those implemented in informal settings (e.g., museums, science centers). A second key charge for the future is to conduct similar research in other disciplines using the techniques of this study. Specifically, we advocate for the use of meta-regression to estimate mean effect size differences across categories while controlling for other influential moderators. Until this work is conducted by the field using comparable techniques, accumulating knowledge about effect size moderation will be challenging.

Until then, our current challenge is to encourage study designers to use the information provided here to more precisely estimate required sample sizes for science education intervention studies. Doing so will decrease the likelihood that study designers will over- or under-recruit schools, teachers, and students, conserving the precious human and financial resources the field needs to continue an agenda of rigorous intervention research.

## Acknowledgments

We gratefully acknowledge assistance of Karen M. Askinas, S. Josh Carson, Lila Goldstein, Jeff Hoover, Brendan Martin, Alexandra Monzon, Jennifer Nichols, Ran Shi, and Jay Wade. We also acknowledge the assistance of Zhipeng Hou for his work on the power calculator. This research was supported by a grant from the National Science Foundation (1118555). The opinions are those of the authors and do not reflect the views of the National Science Foundation.

## Authors' Note

Taylor and Kowalski are co-equal first authors.

## ORCID iDs

J. A. Taylor https://orcid.org/0000-0002-3753-4888
E. Tipton https://orcid.org/0000-0001-5608-1282

## References

*Articles included in the review are marked with an asterisk.*

Afshari, A., & Wetterslev, J. (2015). When may systematic reviews and meta-analyses be considered reliable? *European Journal of Anesthesiology 32*(2). doi:10.1097/EJA.0000000000000186

*Alkhawaldeh, S. A. (2013). Enhancing ninth grade students' understanding of human circulatory system concepts through conceptual change approach. *The European Journal of Social and Behavioural Sciences*, *II*(2), 201–223. doi:10.15405/FutureAcademy/ejsbs(2301-2218).2012.2.7

*Alkhawaldeh, S., & Al Olaimat, A. (2010). The contribution of conceptual change texts accompanied by concept mapping to eleventh-grade students understanding of cellular respiration concepts. *Journal of Science Education and Technology*, *19*(2), 115–125. doi:10.1111/j.1949-8594.2001.tb18010.x

*Alparslan, C., Tekkaya, C., & Geban, O. (2003). Using the conceptual change instruction to improve learning. *Journal of Biological Education*, *37*(3), 135–139. doi:10.1080/00219266.2003.9655868

*Anderson, J. C., II (2007). *Effect of problem-based learning on knowledge acquisition, knowledge retention, and critical thinking ability of agriculture students in urban schools* (Doctoral dissertation, University of Missouri-Columbia). Dissertation

Abstracts International, PhD. (University Microfilms No. 3322674)

*Asay, L. J. (2013). *The importance of explicitly mapping instructional analogies in science education* (Doctoral dissertation, University of Nevada, Las Vegas). Dissertation Abstracts International, PhD. (University Microfilms No. 3590122)

*August, D., Branum-Martin, L., Cardenas-Hagen, 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*, 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

*Azizoglu, N. (2004). *Conceptual change oriented instruction and students' misconceptions in gases*. (Doctoral dissertation, Middle East Technical University). Dissertation Abstracts International, PhD.

*Barthlow, M. J. (2011). *The effectiveness of process oriented guided inquiry learning to reduce alternate conceptions in secondary chemistry* (Doctoral dissertation, Liberty University). Dissertation Abstracts International, Ed.D. (University Microfilms No. 3466432)

*Barthlow, M. J., & Watson, S. B. (2014). The effectiveness of process-oriented guided inquiry learning to reduce alternative conceptions in secondary chemistry. *School Science and Mathematics*, *114*(5), 246–255. doi:10.1111/ssm.12076

*Ben-David, A., & Zohar, A. (2009). Contribution of Meta-strategic knowledge to scientific inquiry learning. *International Journal of Science Education*, *31*(12), 1657–1682. doi:10.1080/09500690802162762

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). New York, NY: Russell Sage.

Bloom, H. S., Bos, J. M., & Lee, S.W. (1999). Using cluster random assignment to measure program impacts. *Evaluation Review*, *23*(4), 445–469. doi:10.1177/0193841X9902300405

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 237–256). New York, NY: Russell Sage.

*Brand Lance, G. (2010). *Evaluating the effects of medical explorers: A case study curriculum on critical thinking, attitude toward life science, and motivational learning strategies in rural high school students* (Doctoral dissertation, Ball State University). Dissertation Abstracts International, Ed.D.

*Cakir, O., Geban, O., & Yuruk, N. (2002). Effectiveness of conceptual change text oriented instruction on students' understanding of cellular respiration concepts. *Biochemistry and Molecular Biology Education*, *30*(4), 239–243. doi:10.1002/bmb.2002.494030040095

*Caleon, I., & Subramaniam, R. (2005). The impact of a cryogenics-based enrichment programme on attitude towards science and the learning of science concepts. *International Journal of Science Education*, *27*(6), 679–704. doi:10.1080/09500690500038306

*Caleon, I. S., & Subramaniam, R. (2007). Augmenting learning in an out-of-school context: The cognitive and affective impact of two cryogenics-based enrichment programmes on upper primary students. *Research in Science Education*, *37*(3), 333–351. doi:10.1007/s11165-006-9032-7

*Cervetti, G. N., Barber, J., Dorph, R., Pearson, P. D., & Goldschmidt, P. G. (2012). The impact of an integrated approach to science and literacy in elementary school classrooms *Journal of Research in Science Teaching*, *49*(5), 631–658. doi:10.1002/tea.21015

*Cetin, G. (2003). *The effect of conceptual change instruction on understanding of ecology concepts* (Doctoral dissertation, the Middle East Technical University). Dissertation Abstracts International, PhD.

*Çetin, P., Kaya, E., & Geban, Ö. (2009). Facilitating conceptual change in gases concepts. *Journal of Science Education and Technology*, *18*(2), 130–137. doi:10.1007/s10956-008-9138-y

*Chang, C. Y., Yeh, T. K., & Barufaldi, J. P. (2010). The positive and negative effects of science concept tests on student conceptual understanding. *International Journal of Science Education*, *32*(2), 265–282. doi:10.1080/09500690802650055

*Chang, H. Y. (2007). *Multilevel-multifaceted approach to assessing the impact of technology-mediated modeling practice on student understanding of the particulate nature of matter* (Doctoral dissertation, University of Michigan). Dissertation Abstracts International, PhD. (University Microfilms No. 3276108)

*Chang, H. Y., & Linn, M. C. (2013). Scaffolding learning from molecular visualizations. *Journal of Research in Science Teaching*, *50*(7), 858–886. doi:10.1002/tea.21089

*Chang, H. Y., Quintana, C., & Krajcik, J. S. (2010). The impact of designing and evaluating molecular animations on how well middle school students understand the particulate nature of matter. *Science Education*, *94*(1), 73–94. doi:10.1002/sce.20352

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*, 283–292. doi:10.3102/0013189X16656615

Cheung, A., Slavin, R. E., Kim, E., & Lake, C. (2016). Effective secondary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching*. doi:10.1002/tea.21338

*Christensen, E .F. (2002). *The effect of homework choices on achievement and intrinsic motivation* (Doctoral dissertation, University of Missouri-St. Louis). Dissertation Abstracts International, Ed.D. (University Microfilms No. 3030104)

*Clark, D., & Jorde, D. (2004). Helping students revise disruptive experientially supported ideas about thermodynamics: Computer visualizations and tactile models. *Journal of Research in Science Teaching*, *41*(1), 1–23. doi:10.1002/tea.10097

*Cobern, W. W., Schuster, D., Adams, B., Applegate, B., Skjold, B., Undreiu, A., & Gobert, J. D. (2010). Experimental comparison of inquiry and direct instruction in science. *Research in Science and Technological Education*, *28*(1), 81–96. doi:10.1080/02635140903513599

*Conklin, E. (2007). *Concept mapping: Impact on content and organization of technical writing in science* (Doctoral dissertation, Walden University). Dissertation Abstracts International, Ed.D. (University Microfilms No. 3254433)

*Cotabish, A., Dailey, D., Hughes, G. D., & Robinson, A. (2011). The effects of a STEM professional development intervention on elementary teachers' science process skills. *Research in the Schools*, *18*(2), 16–25.

*Cotabish, A., Dailey, D., Robinson, A., & Hughes, G. (2013). The effects of a STEM Intervention on Elementary Students' Science Knowledge and Skills. *School Science and Mathematics*, *113*(5), 215–226. doi:10.1111/ssm.12023

*Croom John, R., III (2014). *Knowledge retention for computer simulations: A study comparing virtual and hands-on laboratories* (Doctoral dissertation, Wilkes University, 2013). Dissertation Abstracts International, EdD. (University Microfilms No. 3630293)

*Crowe, J. (2009). *The effect of peer interactions on Newtonian thinking in secondary physics: What are they saying? How does it help?* (Doctoral dissertation, Tufts University). Dissertation Abstracts International, PhD. (University Microfilms No. 3354722)

*Dano, J. B. (2009). *Completing chemistry TAKS Objective 4(9D): The effect of flash animation* (Doctoral dissertation, University of Texas-Pan American). Dissertation Abstracts International, MS. (University Microfilms No. 1468393)

*Davis, E. G. (2007). *a study of the effects of an experimental spiral physics curriculum taught to sixth grade girls and boys* (Doctoral dissertation, Baylor University). Dissertation Abstracts International, Ed.D.

*DeWeese, S. V. (2012). *The effects of mastery learning correctives on academic achievement and student affect* (Doctoral dissertation, Mercer University). Dissertation Abstracts International, PhD. (University Microfilms No. 3528465)

*Ding, N., & Harskamp, E. G. (2011). Collaboration and peer tutoring in chemistry laboratory education. *International Journal of Science Education*, *33*(6), 839–863. doi:10.1080/09500693.2010.498842

*Dogru-Atay, P., & Tekkaya, C. (2008). Promoting students' learning in genetics with the learning cycle. *The Journal of Experimental Education*, *76*(3), 259–280. doi:10.3200/JEXE.76.3.259-280

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold Publishers.

*Ebenezer, J., Chacko, S., Kaya, O. N., Koya, S. K., & Ebenezer, D. L. (2010). Effects of common knowledge construction model sequence of lessons on science achievement and relational conceptual change. *Journal of Research in Science Teaching*, *47*(1), 25–46. doi:10.1002/tea.20295

*Eilam, B. (2002). Strata of comprehending ecology: Looking through the prism of feeding relations. *Science Education*, *86*(5), 645–671. doi:10.1002/sce.10041

*Field, G. B. (2007). *The effect of using Renzulli learning on student achievement: An investigation of internet technology on reading fluency and comprehension*. (Doctoral dissertation, University of Connecticut). Dissertation Abstracts International, PhD.

Fisher, Z., Tipton, E., & Hou, Z. (2017). robumeta: An R-package for robust variance estimation in meta-analysis. arXiv: 1503.02220 [stat.ME]

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.

*Giacalone, V. A. (2004). *Measuring the academic, social, and psychological effects of academic service learning on middle school students* (Doctoral dissertation, Utah State University). Dissertation Abstracts International, PhD. (University Microfilms No. 3122881)

*Gierus, B. J. (2011). *Learning with visual representations through cognitive load theory* (Doctoral dissertation, McGill University, Montreal). Dissertation Abstracts International, MA.

*Granger, E. M., Bevis, T. H., Saka, Y., & Southerland, S. A. (2009). *Comparing the efficacy of reform-based and traditional/verification curricula to support student learning about space science. Paper presented at the meeting of the Annual meeting of the National Association for Research in Science Teaching*, Garden Grove, CA.

*Granger, E. M., Bevis, T. H., Saka, Y., Southerland, S. A, Sampson, V., & Tate, R. L. (2012). The efficacy of student-centered instruction in supporting science learning. *Science*, *338*(105), 1–18. doi:10.1126/science.1223709

*Greenleaf, C., Hanson, T., Herman, J., Litman, C., Madden, S., Rosen, R., & Silver, D. (2009). *Integrating literacy and science instruction in high school biology: Impact on teacher practice, student engagement, and student achievement*. Washington, DC: National Science Foundation.

*Greenleaf, C., Litman, C., Hanson, T., Boscardin, C. K., Herman, J., & Schneider, S. (2011). Integrating literacy in science biology: Teaching and learning impacts of reading apprenticeship professional development. *American Educational Research Journal*, *48*(3), 647–717. doi:10.3102/0002831210384839

*Gulati, S. (2005). *A comparison of inquiry-based teaching through concept maps and traditional teaching in biology* (Doctoral dissertation, University of South Dakota). Dissertation Abstracts International, Ed. D. (University Microfilms No. 3188183)

*Gunel, M., Hand, B., & Gunduz, S. (2006). Comparing student understanding of quantum physics when embedding multimodal representations into two different writing formats: Presentation format versus summary report format. *Science Education*, *90*(6), 1092–1112. doi:10.1002/sce.20160

*Hadzigeorgiou, Y., Anastasiou, L., Konsolas, M., & Prevezanou, B. (2009). A study of the effect of preschool children's participation in sensorimotor activities on their understanding of the mechanical equilibrium of a balance beam. *Research in Science Education*, *39*(1), 39–55. doi:10.1007/s11165-007-9073-6

*Hamilton-Ekeke, J. T. (2007). Relative effectiveness of expository and field trip methods of teaching on students' achievement in ecology. *International Journal of Science Education*, *29*(15), 1869–1889. doi:10.1080/09500690601101664

*Hand, B., Gunel, M., & Ulu, C. (2009). Sequencing embedded multimodal representations in a writing to learn approach to the teaching of electricity. *Journal of Research in Science Teaching*, *46*(3), 225–247. doi:10.1002/tea.20282

*Hand, B., Wallace, C. W., & Yang, E. M. (2004). Using a science writing heuristic to enhance learning outcomes from laboratory activities in seventh-grade science: Quantitative and qualitative aspects. *International Journal of Science Education*, *26*(2), 131–149. doi:10.1080/0950069032000070252

*Haslam, C. Y., & Hamilton, R. J. (2010). Investigating the use of integrated instructions to reduce the cognitive load associated with doing practical work in secondary school science. *International Journal of Science Education*, *32*(13), 1715–1737. doi:10.1080/09500690903183741

Hedges, L. V., & Rhoads, C. (2009). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. doi:10.1002/jrsm.5

*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., & Green, S. (Eds.). *Cochrane handbook for systematic reviews of interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Retrieved from http://www.cochrane-handbook.org.

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

*Howe, C., Devine, A., & Tavares, J. T. (2013). Supporting conceptual change in school science: A possible role for tacit understanding. *International Journal of Science Education*, *35*(5), 864–883. doi:10.1080/09500693.2011.585353

Institute of Education Sciences. (2012). *WWC evidence review protocol for science interventions* [Version 2.0]. Retrieved from http://ies.ed.gov/ncee/wwc/Document/233

IntHout, J., Ioannidis, J., Borm, G., & Goeman, J. (2015). Small studies are more heterogeneous than large ones: A meta-meta-analysis, *Journal of Clinical Epidemiology*, *68*(8). doi:10.1016/j.jclinepi.2015.03.017

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine*, *30*(20), 2481–2498.

*Jang, S. J. (2006). The effects of incorporating Web-assisted learning with team teaching in seventh-grade science classes. *International Journal of Science Education*, *28*(6), 615–632. doi:10.1080/09500690500339753

*Jang, S. J. (2010). The impact on incorporating collaborative concept mapping with coteaching techniques in elementary science classes. *School Science and Mathematics*, *110*(2), 86–97. doi:10.1111/j.1949-8594.2009.00012.x

*Karaçalli, S., & Korur, F. (2014). The effects of project-based learning on students' academic achievement, attitude, and retention of knowledge: The subject of "electricity in our lives." *School Science and Mathematics*, *114*(5), 224–235. doi:10.1111/ssm.12071

*Kay, R., Zucker, A., & Staudt, C. (2012). *Being smart about SmartGraphs: Findings from an experimental trial in physical science classrooms*. Concord, MA: Concord Consortium.

*Khishfe, R. (2012). Nature of science and decision-making. *International Journal of Science Education*, *34*(1), 67–100. doi:10.1080/09500693.2011.559490

*Kiboss, J. K. (2002). Impact of a CBI in physics on students' understanding of measurement concepts and skills associated with school science. *Journal of Science Education and Technology*, *11*, 193–198. doi:10.1023/A:1014673615275

*Kiboss, J., Ndirangu, M., & Wekesa, E. (2004). Effectiveness of a computer-mediated simulations program in school biology on pupils' learning outcomes in cell theory. *Journal of Science Education and Technology*, *13*(2), 207–213. doi:10.1023/B:JOST.0000031259.76872.f1

*Klenk, K. E. (2011). *Computer animation in teaching science: Effectiveness in teaching retrograde motion to 9th graders* (Doctoral dissertation, University of Rhode Island). Dissertation Abstracts International, PhD. (University Microfilms No. 3487740).

Konstantopolous, S. (2008). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *1*, 66–88. doi:10.1080/19345740701692522

*Kopec, R. H. (20002). *Virtual, on-line, frog dissection vs. conventional laboratory dissection: A comparison of student achievement and teacher perceptions among honors, general ability, and foundations-level high school biology classes* (Doctoral dissertation, Seton Hall University, College of Education and Human Services). Dissertation Abstracts International, Ed.D. (University Microfilms No. 3040985)

*Lamb, R. L., & Annetta, L. (2013). The use of online modules and the effect on student outcomes in a high school chemistry class. *Journal of Science Education and Technology*, *22*(5), 603–613. doi:10.1007/s10956-012-9417-5

*Lastica, J. R., & O'Donnell, C. L. (2007). Considering the role of fidelity of implementation in science education research: Fidelity as teacher and student adherence to structure. In O'Donnell, C. L. (Chair), *Analyzing the relationship between Fidelity of Implementation (FOI) and student outcomes in a quasi-experiment*. Paper presented at the meeting of the Annual Meeting of the American Educational Research Association, Chicago, IL.

*Lazarowitz, R., & Naim, R. (2013). Learning the cell structures with three-dimensional models: Students' achievement by methods, type of school and questions' cognitive level. *Journal of Science Education and Technology*, *22*(4), 500–508. doi:10.1007/s10956-012-9409-5

*Lee, M. K., & Erdogan, I. (2007). The effect of science-technology-society teaching on students' attitudes toward science and certain aspects of creativity. *International Journal of Science Education*, *29*(11), 1315–1327. doi:10.1080/09500690600972974

*Lin, H. S., & Chen, C. C. (2002). Promoting preservice chemistry teachers' understanding about the nature of science through history. *Journal of Research in Science Teaching*, *39*(9), 773–792. doi:10.1002/tea.10045

*Lin, H. S., Hong, Z. R., Chen, C. C., & Chou, C. H. (2011). The effect of integrating aesthetic understanding in reflective inquiry activities. *International Journal of Science Education*, *33*(9), 1199–1217. doi:10.1080/09500693.2010.504788

*Lin, H. S., Hung, J. Y., & Hung, S. C. (2002). Using the history of science to promote students' problem-solving ability. *International Journal of Science Education*, *24*(5), 453–464. doi:10.1080/09500690110073991

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

*Lynch, S., Kuipers, J., Pyke, C., & O'Donnell, C. (2007). *Scaling up Curriculum for Achievement, Learning, and Equity Project*

*Annual Report* (Research Rep. No. 14). Washington, DC: George Washington University.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304–316. doi:10.3102/0013189X14545513

*Mandrin, P. A., & Preckel, D. (2009). Effect of similarity-based guided discovery learning on conceptual performance. *School Science and Mathematics*, *109*(3), 133–145. doi:10.1111/j.1949-8594.2009.tb17949.x

*Marbach-Ad, G., Rotbain, Y., & Stavy, R. (2008). Using computer animation and illustration activities to improve high school students' achievement in molecular genetics. *Journal of Research in Science Teaching*, *45*(3), 273–292. doi:10.1002/tea.20222

*Mbajiorgu, N. M., & Ali, A. (2003). Relationship between STS approach, scientific literacy, and achievement in biology. *Science Education*, *87*(1), 31–39. doi:10.1002/sce.10012

*Mbajiorgu, N. M., Ezechi, N. G., & Idoko, E. C. (2007). Addressing nonscientific presuppositions in genetics using a conceptual change strategy. *Science Education*, *91*(3), 419–438. doi:10.1002/sce.20202

*Merrill, C. (2001). Integrated technology, mathematics, and science education: A quasi-experiment. *Journal of Industrial Teacher Education*, *38*(3), 45–61.

*Michalsky, T. (2013). Integrating skills and wills instruction in self-regulated science text reading for secondary students. *International Journal of Science Education*, *35*(11), 1846–1873. doi:10.1080/09500693.2013.805890

*Mikkila-Erdmann, M. (2001). Improving conceptual change concerning photosynthesis through text design. *Learning and Instruction*, *11*, 241–257. doi:10.1016/S0959-4752(00)00041-4

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.

*Mueller, A. L. (2009). *The effects of the Apple Genomics Project active-learning lessons on high school students' knowledge, motivation and perceptions of learning experiences and teachers' perceptions of teaching experiences* (Doctoral dissertation, Purdue University). Dissertation Abstracts International, MS. (University Microfilms No. 1469893).

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford, UK: Oxford University Press.

*Musallam, R. (2010). *The effects of using screencasting as a multimedia pre-training tool to manage the intrinsic cognitive load of chemical equilibrium instruction for advanced high school chemistry students* (Doctoral dissertation, University of San Francisco). Dissertation Abstracts International, Ed.D. (University Microfilms No. 3416991)

*Nigro, R. G., & Trivelato, S. F. (2012). Knowledge, its application, and attitudes associated with the reading of diverse genres of science texts. *International Journal of Science Education*, *34*(16), 2529–2564. doi:10.1080/09500693.2012.711916

*O'Donnell, C. L. (2007). *Fidelity of implementation to instructional strategies as a moderator of curriculum unit effectiveness in a large-scale middle school science quasi-experiment* (Doctoral dissertation, George Washington University). Dissertation Abstracts International, Ed.D. (University Microfilms No. 3276564)

*Olgun, O. S., & Adali, B. (2008). Teaching grade 5 life science with a case study approach. *Journal of Elementary Science Education*, *20*(1), 29–44. doi:10.1007/BF03174701

Polanin, J. R., & Pigott, T. D. (2015). The use of meta-analytic statistical significance testing. *Research Synthesis Methods*, *6*(1), 63–73.

Pustejovsky, J. (2015). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections*. R package version 0.0. 0.9000. Retrieved from https://github.com/jepusto/clubSandwich

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

*Ramseyer, D. L. (2012). *Conceptual change: the integration of geologic time into the teaching of evolution* (Doctoral dissertation, Mississippi State University). Dissertation Abstracts International, PhD. (University Microfilms No. 1530712)

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*(2), 173–185. doi:10.1037/1082-989X.2.2.173

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. doi:10.1037/1082-989X.5.2.199

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29. doi:10.3102/0162373707299460

*Rivard, L. P. (2004). Are language-based activities in science effective for all students, including low achievers? *Science Education*, *88*(3), 420–442. doi:10.1002/sce.10114

*Rivard, L. P., & Straw, S. B. (2000). The effect of talk and writing on learning science: An exploratory study. *Science Education*, *84*(5), 566–593. doi:10.1002/1098-237X(200009)84:5<566::AID-SCE2>3.0.CO;2-U

*Romu, T. (2008). *Demystifying misconceptions in grade 12 electrochemistry* (Doctoral dissertation, University of Manitoba (Canada)). Dissertation Abstracts International, MEd.

*Rosebrock, M. M. (2007). *The effect of systematic vocabulary instruction on the science achievement of fifth-grade students* (Doctoral dissertation, University of Houston). Dissertation Abstracts International, EdD. (University Microfilms No. 3272592)

*Rotbain, Y., Marbach-Ad, G., & Stavy, R. (2006). Effect of bead and illustrations models on high school students' achievement in molecular genetics. *Journal of Research in Science Teaching*, *43*(5), 500–529. doi:10.1002/tea.20144

*Rotbain, Y., Marbach-Ad, G., & Stavy, R. (2008). Using a computer animation to teach high school molecular biology. *Journal of Science Education and Technology*, *17*(1), 49–58. doi:10.1007/s10956-007-9080-4

*Ruby, A. (2006). Improving science achievement at high-poverty urban middle schools. *Science Education*, *90*(6), 1005–1027. doi:10.1002/sce.20167

*Sampson, V., & Clark, D. (2009). The impact of collaboration on the outcomes of scientific argumentation. *Science Education*, *93*(3), 448–484. doi:10.1002/sce.20306

*Scharfenberg, F. J., & Bogner, F. X. (2011). A new two-step approach for hands-on teaching of gene technology: Effects on students' activities during experimentation in an outreach gene technology lab. *Research in Science Education*, *41*(4), 505–523. doi:10.1007/s11165-010-9177-2

*Scharfenberg, F-J., Bogner, F. X., & Klautke, S. (2006). The suitability of external control-groups for empirical control purposes: A cautionary story in science education research; 12th grade. *Electronic Journal of Science Education*, *11*(1), 22–36.

*Scharfenberg, F. J., Bogner, F. X., & Klautke, S. (2007). Learning in a gene technology lab with educational focus: Results of a teaching unit with authentic experiments. *Biochemistry and Molecular Biology Education*, *35*, 28–39. doi:10.1002/bmb.1

*Scharfenberg, F. J., Bogner, F. X., & Klautke, S. (2008). A category-based video analysis of students' activities in an out-of-school hands-on gene technology lesson. *International Journal of Science Education*, *30*(4), 451–467. doi:10.1080/09500690701213898

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62–87. doi:10.3102/1076998607302714

*Schuster, D., Cobern, W. W., Adams, B., Undreiu, A., Skjold, B., & Applegate, B. (2015). *Learning of core disciplinary ideas: Efficacy comparison of two contrasting modes of science instruction*. Unpublished manuscript, the Mallinson Institute for Science Education, Western Michigan University, Kalamazoo, MI.

*Sesen, B. A., & Tarhan, L. (2013). Inquiry-based laboratory activities in electrochemistry: High school students' achievements and attitudes. *Research in Science Education*, *43*(1), 413–435. doi:10.1007/s11165-011-9275-9

Slavin, R. E., Lake, C., Hanley, P., & Thurston, A. (2014). Experimental evaluations of elementary science programs: A best-evidence synthesis. *Journal of Research in Science Teaching*, *51*(7), 870–901. doi:10.1002/tea.21139

*Smith, J. A. R. (2010). *Historical short stories and the nature of science in a high school biology classroom* (Doctoral dissertation, Iowa State University). Dissertation Abstracts International, MS. (University Microfilms No. 1476350).

Spybrook, J., Bloom, H., Congdon, R., Liu, X., Martinez, A., & Raudenbush, S. (2011). Optimal Design Plus Empirical Evidence, v. 3.01 [Computer software].

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1). doi:10.1177/2332858415625975

*Stern, L., Barnea, N., & Shauli, S. (2008). The effect of a computerized simulation on middle school students' understanding of the kinetic molecular theory. *Journal of Science Education and Technology*, *17*(4), 305–315. doi:10.1007/s10956-008-9100-z

*Swaak, J., De Jong, T., & Van Joolingen, W. R. (2004). The effects of discovery learning and expository instruction on the acquisition of definitional and intuitive knowledge. *Journal of Computer Assisted Learning*, *20*(4), 225–234. doi:10.1111/j.1365-2729.2004.00092.x

Tanner-Smith, E. E., Tipton, E., & Polanin, J. D. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental Life Course Criminology*, *2*(1), 85–112. doi:10.1007/s40865-016-0026-5

*Tai, C. C. (2009). *Students' understanding of combustion and its instruction* (Doctoral dissertation, Columbia University). Dissertation Abstracts International, PhD. (University Microfilms No. 3346246)

*Tarhan, L., Ayar-Kayali, H., Urek, R. O., & Acar, B. (2008). Problem-based learning in 9th grade chemistry class: "Intermolecular forces." *Research in Science Education*, *38*(3), 285–300. doi:10.1007/s11165-007-9050-0

*Tastan, Ö., Yalçinkaya, E., & Boz, Y. (2008). Effectiveness of conceptual change text-oriented instruction on students' understanding of energy in chemical reactions. *Journal of Science Education and Technology*, *17*(5), 444–453. doi:10.1007/s10956-008-9113-7

Taylor, J., Furtak, E., Kowalski, S., Martinez, A., Slavin, R., Stuhlsatz, M., & Wilson, C. (2016). Emergent themes from recent research syntheses in science education and their implications for research design, replication, and reporting practices. *Journal of Research in Science Teaching*, *53*(8), 1216–1231. doi:10.1002/tea.21327

Tipton, E. (2013) Robust variance estimation in meta-regression for binary dependent outcomes. *Research Synthesis Methods*, *4*(2), 169–187. doi:10.1002/jrsm.1070

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

Tipton, E., & Pustejovsky, J. (2015) Small-sample adjustments to multivariate hypothesis tests in robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*(6), 604–634. doi:10.3102/1076998615606099

Turner, R. M., Bird, S. M., & Higgins, J. P. T. (2013). The impact of study size on meta-analyses: Examination of underpowered studies in *Cochrane Reviews. PLoS ONE*, *8*(3). doi:10.1371/0059202

*Tyler-Wood, T., Ellison, A., Lim, O., & Periathiruvadi, S. (2012). Bringing Up Girls in Science (BUGS): The effectiveness of an afterschool environmental science program for increasing female students' interest in science careers. *Journal of Science Education and Technology*, *21*(1), 46–55. doi:10.1007/s10956-011-9279-2

*Ülen, S., Branka, C., Slavinec, M., & Gerlic, I. (2014). Designing and evaluating the effectiveness of physlet-based learning materials in supporting conceptual learning in secondary school physics. *Journal of Science Education and Technology*, *23*, 658–667. doi:10.1007/s10956-014-9492-x

*Vaca, J. L., Jr. (2010). *The effect of constructivist teaching strategies on science test scores of middle school students* (Doctoral dissertation, Walden University). Dissertation Abstracts International, Ed.D. (University Microfilms No. 3418952)

*Wecker, C., Rachel, A., Heran-Dörr, E., Waltner, C., Wiesner, H., & Fischer, F. (2013). Presenting theoretical ideas prior to inquiry activities fosters theory-level knowledge. *Journal of Research in Science Teaching*, *50*(10), 1180–1206. doi:10.1002/tea.21106

*Wekesa, E., Kiboss, J., & Ndirangu, M. (2006). Improving students' understanding and perception of cell theory in school biology using a computer-based instruction simulation program (Doctoral dissertation, Egerton University). *Dissertation Abstracts International*, *15*(4), 397–410.

Westine, C. D. (2016). Finding efficiency in the design of large multisite evaluations: Estimating variances for science achievement studies. *American Journal of Evaluation*, *37*(3), 311–325. doi:10.1177/1098214015624014

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490–519. doi:10.1177/0193841X14531584

*What Works Clearinghouse, IES (2012). *Great Explorations in Math and Science® (GEMS®) Space science sequence* (WWC Intervention Report). Washington, DC: Author.

*What Works Clearninghouse (2012). *Astronomy Resources for Intercurricular Elementary Science (ARIES): Exploring motion and forces* (WWC Intervention Report). Washington, DC: Author.

*Wheeler, T. L. (2007). *Effectiveness of an electronic microworld text on student learning in chemistry*. (Doctoral dissertation, Utah State University, 2006). Dissertation Abstracts International, EdD. (University Microfilms No. 3252276).

*Wiggins, F. (2006). *The effects of hands-on-science instruction on the science achievement of middle school students* (Doctoral dissertation, Texas Southern University). Dissertation Abstracts International, EdD. (University Microfilms No. 3317522)

*Yenilmez, A., & Tekkaya, C. (2006). Enhancing students' understanding of photosynthesis and respiration in plants through conceptual change approach. *Journal of Science Education and Technology*, *15*(1), 81–87. doi:10.1007/s10956-006-0358-8

*Yuruk, N. (2007). The effect of supplementing instruction with conceptual change texts on students' conceptions of electrochemical cells. *Journal of Science Education and Technology*, *16*(6), 515–523. doi:10.1007/s10956-007-9076-0

*Zucker, A., Kay, R., & Staudt, C. (2014). Helping students make sense of graphs: An experimental trial of SmartGraphs software. *Journal of Science Education and Technology*, *23*(3), 441–457. doi:10.1007/s10956-013-9475-3

## Authors

JOSEPH A. TAYLOR is principal scientist at BSCS Science Learning. His research focuses on the effectiveness of science education interventions, optimal design of and reporting from intervention studies, and issues around knowledge accumulation from STEM education research.

SUSAN M. KOWALSKI is a senior research scientist at BSCS Science Learning. Her research encompasses two major strands: meta-analyses of science education research and research on curriculum and professional development programs for middle and high school teachers and students.

JOSHUA R. POLANIN is a principal researcher at the American Institutes for Research. Dr. Polanin's research focuses on improving the methods of meta-analysis through pragmatic approaches to data management and analyses.

KAREN ASKINAS is a research associate at BSCS Science Learning. She has developed and maintained large data sets for meta-analyses and for teacher and student assessments in science education research projects.

MOLLY A. M. STUHLSATZ is a research scientist at BSCS Science Learning. Her recent research focuses on the impact of leadership development on district-level student outcomes in science, investigating the effectiveness of videocase-based teacher professional development, and examining the feasibility of using of computer scoring models to score open-ended assessment items.

CHRISTOPHER D. WILSON is a senior research scientist and director of research at BSCS Science Learning. His research focuses on the assessment of teacher and student learning in science education, the impact of lesson analysis based professional development, and the application of automated scoring techniques to the measurement of teacher PCK and student argumentation.

ELIZABETH TIPTON is an assistant professor of applied statistics in the Human Development Department at Teachers College, Columbia University. Her research focuses on the development of methods for improving generalizations from cluster randomized and multi-site experiments—including improved site selection and recruitment strategies—and on methods for estimation of treatment impacts in these studies and through meta-analysis.

SANDRA JO WILSON is a principal associate at Abt Associates, Inc. Her recent research focuses on risk factors associated with school failure, the prevention of high school dropout, parenting and family-based interventions, and the effectiveness of educational interventions and practices for the What Works Clearinghouse.