Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1248

The Effects of Early Childhood Science Educational Interventions on Children's Science Achievement: A Meta-Analysis of Classroom-Based Studies

Kathleen Lynch
University of Connecticut

Catherine Asher
University of Michigan

Amelia Gotwals
Michigan State University

John Settlage
University of Connecticut

The importance of providing children with more robust opportunities to access high-quality science instruction is a widely recognized challenge. Unfortunately, science instruction is often neglected in the earliest school grades, meaning that many young children face opportunity gaps to learning science. We present the results of a meta-analysis of experimental and quasiexperimental research of the effects of classroom-based pre-K-1st grade science educational interventions. We find that, on average, treatment group children in early science educational interventions demonstrated significantly stronger science achievement outcomes compared to control group children. Most of the programmatic and contextual moderator variables examined were not significant predictors of the magnitudes of achievement impacts. Rather, it appeared that a variety of different intervention strategies could effectively improve early science learning outcomes. We found suggestive evidence of positive impacts on children's literacy and socialemotional learning outcomes, and on teacher-level outcomes. We discuss implications and promising directions for future research.

VERSION: July 2025

Suggested citation: Lynch, Kathleen, Catherine Armstrong Asher, Amelia Gotwals, and John Settlage. (2025). The Effects of Early Childhood Science Educational Interventions on Children's Science Achievement: A Meta-Analysis of Classroom-Based Studies. (EdWorkingPaper: 25-1248). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/drc7-7b08

The Effects of Early Childhood Science Educational Interventions on Children's Science Achievement: A Meta-Analysis of Classroom-Based Studies

Kathleen Lynch[1], Catherine Asher[2], Amelia Gotwals[3], John Settlage[1]
[1]University of Connecticut
[2]University of Michigan
[3]Michigan State University

April 2025

Abstract

The importance of providing children with more robust opportunities to access high-quality science instruction is a widely recognized challenge. Unfortunately, science instruction is often neglected in the earliest school grades, meaning that many young children face opportunity gaps to learning science. We present the results of a meta-analysis of experimental and quasi-experimental research of the effects of classroom-based pre-K-1st grade science educational interventions. We find that, on average, treatment group children in early science educational interventions demonstrated significantly stronger science achievement outcomes compared to control group children. Most of the programmatic and contextual moderator variables examined were not significant predictors of the magnitudes of achievement impacts. Rather, it appeared that a variety of different intervention strategies could effectively improve early science learning outcomes. We found suggestive evidence of positive impacts on children's literacy and social-emotional learning outcomes, and on teacher-level outcomes. We discuss implications and promising directions for future research.

Keywords: Science education; early childhood; meta-analysis; professional development

Author Note

Correspondence concerning this article should be addressed to Dr. Kathleen Lynch, kathleen_lynch@uconn.edu.

The Effects of Early Childhood Science Educational Interventions on Children's Science Achievement: A Meta-Analysis of Classroom-Based Studies

The need to provide children with more opportunities to access high-quality science instruction that strengthens their learning is a widely recognized challenge facing education (National Academies of Sciences, Engineering, and Medicine [NASEM], 2022; National Research Council [NRC], 2007). Evidence suggests that early exposure to standards-aligned science instruction can have short- and long-term impacts on students' science achievement (Curran & Kitchin, 2019; Gropen et al., 2017; Kaderavek et al., 2020). Early science instruction also has the potential to foster broader skills in domains such as early literacy (e.g., Paprzycki et al., 2017; Raven & Wenner, 2022) and social-emotional learning (SEL) (e.g., Bustamante et al., 2018), and could potentially lay an early foundation for children's development of the science-related knowledge and skills valuable for informed citizenship (NRC, 2012).

Young children, whom we hereafter define for the purposes of this article to include children in grades pre-K-1, have strong levels of curiosity and interest in science[1], and bring with them to school rich funds of science experiences and knowledge cultivated in their homes and communities (e.g., Dabney et al., 2013; Edwards et al., 2016). Yet despite these strengths, in practice children often do not have adequate opportunities to develop their science proficiency in early grades classrooms. As of fourth grade, 64% of U.S. children overall failed to meet science proficiency benchmarks on the 2019 NAEP; among low-income children, the rate was 80% (U.S. Department of Education, 2019). Early grades teachers often face pressures to raise mathematics and reading scores, and these pressures frequently come at a cost in terms of instructional time and professional development (PD) for science (Anderson, 2011; Early et al., 2010; Judson,

[1] Following others (e.g., Bustamante et al., 2018), we use the term 'science' for brevity throughout the article when referring to early childhood interventions in both science and/or engineering.

2013; Plumley, 2019). Further, early grades teachers tend to be generalists in their academic content training (NASEM, 2022) and to have little preparation specific to science instruction. These issues may be exacerbated in high-poverty settings, where accountability demands are often especially high, and schools frequently face constraints in securing instructional resources (Banilower et al., 2018; Ladson-Billings, 2006). Given these various factors, concerns remain that children's potential for science learning in early grades classrooms is not being fully realized (NASEM, 2025).

In recent years, researchers, policymakers, and funders have turned growing attention to science education starting in pre-K. These changes reflect both growing recognition of the importance of the early childhood period for children's cognitive and affective development generally (Institute of Medicine and National Research Council, 2000) and increasing awareness that young children are able to develop foundational science knowledge and practices that can strengthen their ability to comprehend more advanced content in the later grades (e.g., Bustamante et al., 2018; Ravanis, 2017; Trundle & Saçkes, 2015). While the Next Generation Science Standards (NGSS Lead States, 2013) and *Framework for K-12 Science Education* (NRC, 2012) were written to apply to grades K-12, and did not encompass preschool/pre-K, important research and policy documents have since discussed how core instructional concepts addressed in these documents could be translated to the pre-K level (e.g., NASEM, 2022; 2023) and have further elaborated on specific considerations for standards-aligned science teaching in the pre-K through early elementary grades (NASEM, 2022). Frameworks and standards for science learning have been adopted by numerous state pre-K programs (Friedman-Krauss et al., 2022). Additionally, the Head Start program, the major U.S. federal preschool program that serves children from low-income backgrounds, has defined a developmental progression and learning indicators for preschoolers in scientific reasoning (U.S. Department of Health & Human Services, 2015).

This interest in young children's science learning has also catalyzed numerous new research studies evaluating the impacts of innovative early childhood science curricula and classroom interventions on student learning. However, these programs may vary in their efficacy.

In the current study, we conducted a comprehensive meta-analysis to examine the impacts of early childhood science educational interventions on children's science achievement. We also examined effects on children's literacy and SEL outcomes, which are beneficial both for future performance in science coursework and for academic and career attainment (NRC, 2014), and on teacher-level outcomes, which are critical for understanding how interventions achieve their effects. We further sought to identify factors that predict variations in interventions' impacts on children's science learning.

Prior Research Synthesizing the Results of Science Classroom Interventions

Several informative prior studies have reviewed the empirical literature on the impacts of science classroom interventions. For example, evidence from several reviews points to the value of active learning in PD as well as specific pedagogical tools (e.g., questioning, assessment, instructional technology, and collaborative learning) for bolstering programmatic outcomes in science education (Savelsbergh et al., 2016; Schroeder et al., 2007; You et al., 2024). Some of these conclusions are limited by reviewers' inclusion of pre-post studies, which generate biased findings because such designs cannot distinguish intervention-induced growth from score increases associated with the passage of time as children mature (Campbell, 1957). Regarding measurement, Taylor et al. (2018) identified a link between science interventions' impacts and outcome measure type, finding that impacts tended to be larger on researcher-developed assessments as compared to standardized assessments, a common finding in the education literature more generally (e.g., Hill et al., 2008).

Because these reviews relied primarily on studies conducted with older students, this limits the translation of results to science classroom interventions serving the youngest learners. To illustrate this problem, Slavin et al.'s (2014) review of elementary science interventions' impacts on students' learning contained only two studies conducted with children who were first graders and younger. In related reviews, Lynch et al. (2019) found only seven studies within this age range, You et al. (2024) found two, and Estrella et al. (2018) found two. Other useful reviews about teacher coaching (Kraft et al., 2018), mathematics and science teaching (Savelsbergh et al., 2016), inquiry-based science instructional methods (Furtak et al., 2012), and teacher professional development programs (Kennedy, 2016) contained no studies conducted with students below grade 2. Hence, much more could be understood about the efficacy of early childhood science educational interventions, and the factors more strongly producing positive impacts on learning, via a review specifically and exclusively focused on the youngest learners.

Focusing on the Efficacy of Classroom-Based Educational Interventions for Science Learners at the Youngest Grade Levels

It is unclear whether we should expect the same characteristics found to predict positive science outcomes for older students to also apply for young children. Commonly, participants and features of science educational interventions such as the students, their teachers, the materials, and the assessments are expected to differ in important ways between the early grades versus older grade levels. For example, compared to older students, young children in grades pre-K-1 are more likely to be pre-readers or early readers, which means learning science content cannot rely on text reading. Science assessments for young children are less likely to consist of written, group standardized tests, and more likely to comprise task-based assessments, scored interviews, and assessments administered one-on-one (Greenfield, 2015). Teachers at the youngest grade levels also tend to have different kinds of training and preparation compared to their counterparts teaching in the upper grades (NASEM, 2022), and specifically, they may have been provided with different forms of science education-related professional learning experiences compared to those provided to teachers of older students. Given these various factors, it is important to know whether particular features of science educational interventions may predict better outcomes specifically in early childhood settings.

Method

Defining the Studies of Interest

We defined the interventions of interest for the current study as educational programs conducted with teachers and children in classroom settings that were meant to bolster children's science and/or engineering learning. The interventions that we identified in the literature all aimed to achieve this goal via one of the following active ingredients: (1) the implementation of PD for teachers, (2) the introduction of new science-related curriculum materials to classroom, or (3) both of these components.[2]

We applied the following criteria to determine studies' inclusion in the meta-analysis. First, interventions must have been conducted in classroom settings. This definition excluded lab-based studies, as well as interventions conducted in camps or other non-classroom settings.

[2] We note that we refer to the interventions we are studying in this meta-analysis as classroom-based educational interventions, not 'PD interventions' or 'curriculum interventions,' because while all interventions aimed to change the classroom instruction that children experienced in early childhood science, not all interventions contained PD, nor did all interventions involve the introduction of new curriculum materials. In particular, three studies' active ingredients did not include PD for teachers (i.e., curriculum-only studies), while six studies' active ingredients did not include the introduction of new curriculum materials for use in the classroom (i.e., PD-only studies). We use the terms 'interventions' and 'programs' interchangeably in this article.

Second, we required that the samples include students and/or teachers at the preschool/pre-K, kindergarten, or first grade levels. This grade span was aligned with the samples employed in a number of the early science intervention studies we identified in the literature. Third, included studies had to compare a treatment group's performance against that of a control/comparison group that did not receive the intervention. Fourth, the study had to supply adequate empirical information to compute at least one effect size. Fifth, the study had to be a classroom-level intervention delivered by classroom teachers. Interventions that were led by individuals other than classroom teachers (e.g., research staff) were excluded. The study must also have been published in 2002 or later. This search period is similar to that of the What Works Clearinghouse (WWC), which typically does not review studies that are more than two decades old due to their lack of applicability to contemporary educational conditions (WWC, 2014). Lastly, the study must have evaluated the impacts of an educational program; studies that only examined the manipulation of a variable (e.g., brief psychology studies on variables' effects; see Slavin, 2008) were excluded.

Literature Search

We conducted a systematic search for both published and unpublished (i.e., grey literature) studies via the following channels. We began by performing electronic searches of the databases Academic Search Premier, EconLit, ERIC, Proquest Dissertations & Theses, and PsychInfo. To develop search keywords, we began with those used in a prior meta-analysis on STEM educational interventions (Lynch et al., 2019), adapting the keywords to reflect a specific focus on early childhood science, and drawing on prior research to develop these adaptations (e.g., Estrella et al., 2018; Garrett et al., 2019; Taylor et al., 2018). We next hand-searched the reference lists of prior reviews. We also searched the abstracts of grant awards made by IES and

NSF relating to early childhood science education, and conducted web searches for publications and presentations resulting from these awards.[3] We also conducted targeted web searches of the websites of the WWC, Regional Education Laboratories, NBER, MDRC, AIR, RAND, Harvard Center on the Developing Child, Mathematica, and the Society for Research on Educational Effectiveness (SREE), and additionally searched Google Scholar for further studies conducted by authors active in the field. After the removal of duplicates, these searches yielded 3,991 records.

Screening Procedures

We conducted study screening in two rounds. In Round 1, for each record identified in the literature search, two screeners reviewed the title and abstract for basic relevance criteria. Many off-topic studies, for example, studies conducted at older grade levels or in lab settings, were excluded at this stage. A total of 171 studies passed this screening round and moved forward to a full-text screen. In Round 2, we read the full text of each of these studies to determine whether the full inclusion criteria were met, including the inclusion of data to permit calculating one or more effect sizes. Once we identified at least one exclusion reason for a study, the study was discarded. A total of 33 reports containing information on 31 studies, 29 of which included student-level outcome measures, passed Round 2 screening and were included in the meta-analysis. A PRISMA diagram documenting the flow of studies is shown in Figure 1.

Potential Moderators of Early Science Educational Interventions' Efficacy

In order to develop analytic codes for meta-analytic moderator analysis, we began with a codebook developed by Lynch et al. (2025) for their review on STEM classroom interventions, adapting it and adding new codes based on existing early childhood and science intervention-specific literature (e.g., NASEM, 2022; Taylor et al., 2018; Xie et al., 2015) to reflect the current

[3] We note that searching grant abstracts is not commonly done in educational meta-analyses. However, we searched these sources as a supplement to the primary grey literature streams we examined (following Lynch et al, 2019).

review's specific focus on early childhood science. For our final codebook, we identified a parsimonious set of potential moderators of programmatic impact for which to conduct moderator analyses. We describe the variables included in the analysis below. See Tables 1 and 2 for descriptive information about the study moderators.

Methodological, Publication, and Contextual Features

The first set of moderators relate to studies' *methodological and publication features.*

Publication Type. First, we coded for publication type, denoting whether the article was peer-reviewed. Effect sizes may be expected to be larger among peer-reviewed studies as compared with studies that were not peer-reviewed, for example, if journal editors or reviewers display a bias in favor of larger effect sizes in the peer review process (Cooper et al., 2019).

Study Methodology. We classified studies by whether they employed a randomized experimental design or a quasi-experimental methodology. Prior research indicates that effect sizes in education research tend to be smaller in randomized trials than in non-RCTs (e.g., Cheung & Slavin, 2016).

Attrition. We also coded studies for whether they reported differential attrition (i.e., of greater than 10%) between the treatment and control groups (WWC, 2022), given its potential to affect validity in experimental research.

Assessment Type. We coded for assessment type by classifying each measure as standardized or researcher-developed. Effect sizes may be expected to be larger among outcomes measured with researcher-developed assessments, as they are generally more directly aligned with the program being investigated (e.g., Hill et al., 2008; Kraft, 2020).

We further coded for three variables relating to studies' *settings and samples*.

Grade Level. First, given that the efficacy of educational interventions may vary by the grade level of students served (Kraft, 2020), we captured the grade level of the sample, operationalized using a dichotomous indicator indexing whether outcomes were measured for students in preschool/pre-K versus for those in kindergarten or grade 1.

Setting Urbanicity. We coded for whether studies were conducted in urban settings or in suburban, rural, or mixed settings, reflecting the fact that the geographical context of schools could influence the kinds of resources available for implementing interventions (WWC, 2021).

Poverty. Lastly, setting poverty level can also affect the conditions and resources in place to support interventions' implementation (WWC, 2021). We thus coded for whether the intervention was implemented in a high-poverty setting, classified using a dichotomous indicator indexing whether a majority of students (50% or more) in the sample were categorized as eligible for a free or reduced price school lunch or otherwise defined as low-income.

Programmatic Characteristics

The second category of moderators comprise programmatic characteristics. Of these, the first set are *programmatic foci of the interventions.*

We began by examining several moderator variables aligned with core facets of the NRC (2012) *Framework. The Framework* presents a vision of science proficiency as emanating from learning that occurs at the nexus of the three dimensions of (1) science and engineering practices [SEPs], (2) crosscutting concepts, and (3) disciplinary core ideas (NRC, 2012; see also NASEM et al., 2023). The practices dimension is discussed below. The crosscutting concepts are "concepts that bridge disciplinary boundaries, having explanatory value throughout much of science and engineering" (NRC, 2012, p. 83). The *Framework* identifies seven crosscutting concepts: (1) patterns; (2) cause and effect: mechanism and explanation; (3) scale, proportion, and quantity; (4) systems and system models; (5) energy and matter: flows, cycles, and conservation; (6) structure and function; and (7) stability and change. Finally, the disciplinary core ideas are classified into physical sciences, life sciences, earth and space sciences, and engineering, technology, and applications of science.

We first examined the potential moderating roles of interventions' foci on practices and crosscutting concepts separately, then explored the role of a joint focus on all three dimensions.

Focal Activities and Practices for Children. We note that the NRC (2012) *Framework* and the NGSS (NGSS Lead States, 2013) cover grades K-12 yet were not explicitly written to apply to pre-K. This is most relevant for the 'practices' dimension (NASEM, 2022), as the forms of science learning practices in which the youngest learners can productively engage may be expected to vary from those for significantly older students. Because a number of studies in our sample included pre-K children, we addressed this by coding for interventions' focus on practices in two different ways (described below). To preview the findings, regardless of which of these variables is used in the analysis, the results are similar.

First, as our primary measure of practices, we categorized interventions by the breadth with which the intervention addressed science- and engineering-related activities of investigation and design for young learners, as defined by NASEM (2022). Specifically, NASEM (2022) identified five forms of activity of investigation and design in which children of preschool through elementary age can beneficially engage which are connected to the kinds of work that scientists and engineers do and the science and engineering practices (SEPs) identified in the *Framework* (NRC, 2012). These forms include (a) orienting to phenomena and design challenges; (b) gathering and analyzing data and information; (c) constructing explanations and design solutions; (d) communicating reasoning to self and others; and (e) connecting learning across content areas and contexts (NASEM, 2022). Each intervention was coded (on a scale from

0-5) for the number of these forms of activity on which the program focused.

Secondly, as a secondary measure of practices and as a sensitivity check, we categorized interventions by the breadth with which the program addressed the science and engineering practices (SEPs) defined in the NRC (2012) *Framework.* As defined in the *Framework*, the SEPs are types of activities that scientists and engineers engage in as they investigate and make sense of the natural and designed world. The *Framework* outlines eight practices: (1) asking questions and defining problems; (2) developing and using models; (3) planning and carrying out investigations; (4) analyzing and interpreting data; (5) mathematics and computational thinking, (6) constructing explanations and defining solutions; (7) engaging in argument from evidence, and (8) obtaining, evaluating, and communicating information. Each intervention was coded (on a scale from 0-8) for the number of SEPs on which the program focused.

Crosscutting Concepts. Next, to capture interventions' focus on crosscutting concepts, we categorized interventions according to whether or not they included a focus on at least one of the seven crosscutting concepts identified in the *Framework* (listed above) (NRC, 2012). This focus could be either explicitly stated or evident from the program elements.

Content Literacy. Additionally, we categorized interventions by whether they focused on developing students' content literacy. We operationalized content literacy interventions as those that aim to integrate domain content knowledge development (here, in science) into literacy instruction (e.g., NRC, 2014; Snow et al., 1998). Such interventions have garnered research attention due to their potential to bolster children's acquiring of skills in core domains of reading and vocabulary while simultaneously helping students connect newly learned science concepts with their existing knowledge base, hence building comprehension of academic content in science.

Duration of Curriculum Exposure. Programs also vary in the amount of time that children were intended to be exposed to the program's curricular activities. Duration of curriculum exposure may matter given the range of evidence suggesting that time on task is related to content learning (Stronge et al., 2011). We began by computing, for each program, the total minutes that children were intended to be exposed to the program curriculum. We then created a dichotomous indicator indexing whether children's intended dosage of the program curriculum was above the median dosage in the study sample (25 hours). As a sensitivity check, models were refit using the continuous dosage hours indicator; results are similar.

We also examined three moderators relating to *features of the professional development* provided to teachers (if any) as part of the intervention.

PD Duration. First, among programs that provided teacher PD, the duration of the PD experiences offered to teachers varies in terms of contact hours. Longer PD duration could translate into more 'time on task' for teacher learning about early childhood science content, pedagogical practices, and curricula (e.g., Desimone, 2009), which could in turn catalyze richer learning experiences offered to students. On the other hand, lengthy PD duration could conceivably be unnecessary or even counterproductive, for example if extra time is not spent in the service of improving science instruction (Garrett et al., 2019; Kennedy, 2016).

PD Focus on Content Knowledge and/or Pedagogical Content Knowledge. Second, PD may or may not include a focus on enriching aspects of teachers' science content knowledge and/or pedagogical content knowledge. Because pre-service education for early grades teachers tends to focus heavily on literacy and numeracy skills (Gotwals et al., 2025; NASEM, 2022), providing teachers with more opportunities to deepen facets of their science-related knowledge may bring affordances as this enhanced knowledge is activated in instruction (NASEM, 2007).

Coaching and/or Mentoring of Teachers. Third, teacher PD may or may not include coaching or mentoring of teachers. Coaches can potentially strengthen teachers' uptake of instructional interventions by observing instruction, providing feedback, and/or helping teachers to troubleshoot intervention implementation issues as they arise (Desimone & Pak, 2017; Gibbons & Cobb, 2017). Prior research at the K-12 level suggests that instructional programs that included coaching as a core element of the PD model tended to have stronger impacts on student achievement compared with those that lacked this element (Kraft et al., 2018), albeit very little of this research has been conducted in early grades science.

Lastly, we conducted one non-pre-specified exploratory analysis. This analysis aimed to explore whether interventions that *attended to the three-dimensional nature of science learning* (NRC, 2012) tended to have larger effects on children's science achievement outcomes compared to interventions that did not have this feature. To operationalize this focus, we categorized each intervention using a dichotomous variable that indexed whether the intervention included foci on all of the three dimensions of (1) NASEM activities (i.e., as our age-specific indicator of practices); (2) crosscutting concepts; and (3) disciplinary core ideas. We then refit the moderation models including this variable.

Coding Procedures

Coding was conducted by the study authors, all of whom hold doctorates in education and have expertise in science education and/or evaluation and research methodology. We first underwent a calibration procedure, coding studies independently. Once all authors reached an inter-rater agreement threshold of above 80% agreement, all studies were double-coded by two study authors working independently, who then reconciled all discrepancies via discussion.

Analysis

We calculated Hedges's *g* effect sizes for each impact estimate reported in the included studies, applying the small-sample bias correction where necessary. Where available, we used model-based estimates of the standardized mean difference as reported by study authors. Otherwise, we calculated standardized mean differences using available test statistics and/or group means and standard deviations.

Meta-Analytic Modeling Approach

We fit correlated and hierarchical effects (CHE) meta-regression models using the *metafor* package in R (Viechtbauer, 2010). This approach accounts for the nesting of multiple effect sizes within each individual study and represents an advance over earlier robust variance estimation (RVE) approaches in that it can accommodate meta-analytic data in the presence of both correlated and hierarchical effects (Pustejovsky & Tipton, 2022). The CHE model assumes that these multiple effect sizes are correlated at a constant rate (rho) that we specified as 0.80 in primary models, following Tanner-Smith and Tipton (2014); we also test alternative values of rho. These models also allow for the inclusion of covariates, which allow us to test for moderating effects of research and intervention features. We report robust standard errors as recommended in Harrer et al. (2021).

Our primary models are as follows. First, to examine overall pooled mean impacts of the interventions, for each category of outcome variables (i.e., all student outcomes; all student achievement outcomes; science achievement outcomes; literacy outcomes; SEL outcomes; and teacher-level outcomes), we fit two models: one unconditional CHE model without covariates, and a second CHE model including a control for study design. Next, to examine potential moderators of interventions' impacts on children's science learning, following our pre-analysis plan, we fit a series of models examining the relationship between each individual moderator and student science achievement outcomes. Lastly, we fit an omnibus model examining the joint relationships between all moderators and student science achievement outcomes.

Results

Study and Intervention Features

In Table 1, we describe the features of the 31 studies included in this meta-analysis and the samples of students they represent. Twenty-nine of the studies include student-level outcome data, yielding a total of 173 impact estimates on student outcomes across multiple domains. The remaining two studies include teacher-level outcomes only.

The majority of studies (71%) were peer-reviewed journal articles, and 19% were conference presentations. Two studies were technical reports, and one was a book chapter. Additionally, 77% of studies used randomized experimental designs, while the remaining 23% employed quasi-experimental methodologies. Over half of the included studies (58%) were conducted in urban settings; the remaining studies were conducted in suburban, rural, or a mix of settings (29%) or did not provide sufficient information to categorize their urbanicity (13%). Finally, 77% of the included studies were conducted with majority low-income samples, among them 23% of studies that included Head Start preschool/pre-K samples.

Table 1 also includes descriptive information on the 173 included student-level effect sizes. Just over half of effect sizes (53%) were derived from standardized assessments, and the remaining 47% were derived from researcher-developed measures. Additionally, 54% of the estimates were drawn from outcomes that assessed pre-K students; 26% came from outcomes that assessed kindergarteners, and 25% came from outcomes assessing first graders (note that students of multiple grade levels completed some assessments).

We report on programmatic features of the interventions in Table 2. All included interventions included science content, in line with the focus of the meta-analysis. Thirteen of the included interventions also incorporated subject matter content in reading/ELA, while 6 studies incorporated specific engineering content and 5 studies included subject matter content in mathematics. Intervention content was coded as including a focus on disciplinary core ideas relating to the domains of life science (k = 23 studies), Earth and space science (k = 13), physical science (k = 20), and/or engineering, technology, and the applications of science (k = 6).

Table 2 also presents information on the potential moderating characteristics of the interventions. Ten studies (32%) included a focus on content-area literacy. The median intended duration of exposure to new intervention curriculum materials for the treatment group was 25 hours (mean: 41 hours). Additionally, in the included interventions, treatment teachers were offered a median of 17 hours of program-related professional development (mean: 53 hours). More than half (61%) of the studies included a focus on at least one of the seven crosscutting concepts identified in the *Framework for K-12 Science Education* (NRC, 2012), whether explicitly stated or evident from the program elements. The studies focus on an average of 3.2 of the five forms of activity of investigation and design identified in NASEM (2022). Coaching or mentoring for teachers was provided in 42% of interventions, while 26% of interventions included PD that contained an explicit focus on deepening aspects of teachers' science and/or engineering-related content knowledge and/or pedagogical content knowledge.

Overall Effects of Early Science Educational Interventions

We find that on average, classroom-based educational interventions focused on early childhood science produced positive pooled mean effects on student outcomes. Table 3 presents the main analytic results for several different sets of effect sizes. First, we consider all 173 student-level effect sizes from the 29 studies with student outcome data, pooled across impact domains. Then, we fit models to examine interventions' impacts specifically on student achievement (using 142 of these effect sizes). Finally, we fit models examining interventions' impacts on science achievement and on literacy achievement separately. Twenty-eight studies contributed 75 science achievement effect sizes, and twelve studies provided 46 reading achievement effect sizes. For each set of outcomes reported, we present results from two different models. The first model is unconstrained, and pools results across all effect sizes and research designs in the relevant category. The second model includes a control for study research design, so that the intercept represents the pooled mean impact for randomized controlled trials.

Considering all student outcome domains together (Table 3, first two columns), we find that classroom-based science interventions increased student outcomes by an average of 0.25 standard deviations ($p < .001$). As shown in the model controlling for research design, the pooled mean impact among RCTs is 0.13 SD ($p < .001$). Table 3 (second two columns) shows the results when restricting the analysis to student achievement outcomes only. Among achievement outcomes, the pooled mean impact is 0.27 SD ($p < .001$) overall, and 0.14 SD ($p < .001$) for RCTs. Lastly, the final two sets of models in Table 3 show impacts on science achievement and on literacy achievement, respectively. We find a pooled mean impact on science achievement of 0.33 SD ($p < .001$) overall, and 0.19 SD ($p < .001$) among RCTs. Meanwhile, among studies that also measured impacts on literacy outcomes, we found a pooled mean impact on literacy achievement of 0.17 SD and 0.14 SD in the unconstrained model and the model controlling for study design, respectively; both of these effects are marginally significant ($p < .10$).

Moderation Results: Study Methodological, Publication, and Contextual Factors

In Table 4, we show the results of fitting a series of models that examine the extent to which effect size magnitudes for science achievement outcomes were associated with study methodological, publication, or coded contextual factors.

While we find that, on average based on the unconstrained models, effect size magnitudes tended to be larger among peer-reviewed studies and for outcomes that were measured with researcher-developed (as opposed to standardized) assessments, these differences were only marginally significant. Once we controlled for whether the study was an RCT, the only difference that remained marginally significant was that for assessment type. None of the other study methodological, contextual, or publication features that we examined, including participant grade level, attrition level, setting urbanicity, or sample poverty level, were significantly associated with effect size magnitudes in any of the models.

Moderation Results: Programmatic Moderators

Table 5 displays the results of fitting a series of CHE models investigating the extent to which interventions' programmatic features moderated estimated impacts on science achievement. We first fit a series of individual models that examined the association between interventions' inclusion of each programmatic feature and effect size magnitudes separately. All models control for research design, so the intercepts represent the pooled mean effect among RCTs without the associated feature. We then fit an omnibus model that investigated all programmatic moderators jointly, allowing us to estimate the association between the presence of each feature and effect size magnitudes, controlling for the other features.

Among the individual models, we found that programs that provided teacher PD of a duration longer than the sample median (i.e., 17 hours) tended to have smaller impacts on children's science achievement (a difference of 0.20 SD; $p < .05$), as compared with programs that provided shorter-duration or no PD. However, this association was not significant in the omnibus model. As a sensitivity test, we refit these models using a continuous indicator for hours of PD duration (see SOM Table S2); this indicator was not significant in either model.

Among the remaining programmatic variables for which we coded, none were significantly associated with effect size magnitudes, for either the individual or omnibus models. More specifically, we did not find significant differences in interventions' pooled mean impacts by the programs' presence of a focus on crosscutting concepts, science/engineering investigation and design activities, content literacy focus, the inclusion of coaching/mentoring, emphasis within the PD on subject matter and/or pedagogical content knowledge development in teachers, nor students' intended exposure time to the program's curricular activities.

Can We Detect Impacts on Teacher Outcomes and/or Children's SEL Outcomes?

While our primary focus was on academic achievement, our dataset also allowed us to explore impacts on teachers' outcomes and on children's SEL outcomes. Eight studies, all RCTs, contributed 44 teacher outcome effect sizes, and eight studies contributed 28 effect sizes for child-level SEL outcomes (e.g., social skills, academic motivation, and persistence).

These results are shown in Table 6. On average, interventions improved teacher outcomes by 0.89 SD ($p < .001$). A subset of these estimates ($n = 30$ effect sizes) were observational measures of classroom instruction (as opposed to beliefs, attitudes, or knowledge). We found a pooled mean impact estimate of 0.94 SD ($p < .001$) on classroom instruction. Other categories of teacher outcome variables contained too few effect sizes to permit separate analyses.

The second set of models shows impacts on students' SEL outcomes. Pooled mean impacts on SEL outcomes are positive in sign (0.24 SD, unconstrained model; 0.12 SD, controlled model); the pooled effect was statistically significant ($p < .05$) in the unconstrained model only. However, we note that the number of studies that contributed outcomes in this category is small; hence power to detect significant effects is limited.

Additional Sensitivity Checks and Publication Bias Analysis

We conducted several analyses to test the sensitivity of the findings to the inclusion of individual studies, analytic decisions, and potential publication bias.

First, we conducted two sensitivity checks relating to the moderator analyses. As described above, we refit all moderation models using the number of SEPs rather than the number of activities of investigation and design. The pattern of findings is similar. Next, as an exploratory analysis, we categorized each intervention using a dichotomous variable that indexed whether the intervention attended to the three-dimensional nature of science learning by including foci on all of the three dimensions of NASEM activities; crosscutting concepts, and disciplinary core ideas. We then refit the moderation models including this variable. The point estimates in both the bivariate and omnibus models are positive but not statistically significant (see SOM Table S3). In sum, these checks did not alter our primary conclusions.

Second, to check that no individual study was driving the identified positive effects, we conducted a leave-one-out robustness check (similar to StataCorp, 2021), in which we sequentially dropped each individual study in the sample, and then re-analyzed the data for science achievement effects using the CHE model. In all iterations of this check, the pooled mean science achievement impact estimate was positive, statistically significant, and between 0.23 and 0.35 SD, and between 0.17 and 0.21 SD when controlling for research design. Next, because CHE models require specifying the assumed correlation coefficient between effect sizes from the same study, we tested the sensitivity of the primary findings to alternate values of the correlation coefficient. As shown in Table SOM S4, across a range of values between 0.2 and 0.9, the pooled mean effect on science achievement, among RCTs, is consistently between 0.19 and 0.21 standard deviations ($p < .001$).

Finally, we conducted several checks to assess the extent to which our findings might have been influenced by publication bias. As we showed in Table 4, we fit models to examine whether peer-reviewed studies tended to show larger effects compared to non-peer-reviewed studies. We found that there was not a statistically significant difference in pooled mean effects by peer review status after controlling for study methodological design.

Second, we conducted a 'trim and fill' analysis (Duval & Tweedie, 2000) in order to test for the potential presence of missing studies from the meta-analytic dataset which would have been expected to have been observed under the assumption of a symmetrical distribution of effects. We conducted trim and fill analyses separately at both the study level and the effect size level. At the effect size level, the analysis identified no 'missing' results that warranted imputation. At the study level, we identified only one 'missing' study-level aggregate effect among the SEL outcome sample; imputing this value does not change the pooled mean SEL outcome estimate. No 'missing' study-level aggregate effects were identified for science achievement outcomes, our primary outcome domain of interest.

Lastly, Egger's test is a regression-based approach that further investigates whether there is sufficient evidence to suggest the presence of publication bias (Egger et al., 1997). Because Egger's test has inflated Type I errors with continuous outcomes and in the presence of hierarchical effects (Pustejovsky & Rodgers, 2019; Rodgers & Pustejovsky, 2020), we conduct two modified versions of the Egger's test to test for asymmetry in the distribution of effects on science achievement (see SOM Table S5). First, we aggregate results to the study level and regress the study average normal standard deviation, i.e., the mean effect size divided by the mean standard error, on the inverse of the study mean standard error. The results indicate no evidence of publication bias. Second, we include the standard error as a moderator in a meta-analytic model, implemented with robust variance estimation as recommended by Rodgers and Pustejovsky (2020). Again, we fail to reject the null hypothesis at the $p < .05$ level. Taken together, the completed tests show little indication that the primary findings are being driven by publication bias, supporting the robustness of our summary conclusions.

Discussion

Overall Effects of Classroom Interventions Focused on Early Childhood Science

In sum, we found that early childhood science educational interventions of the types examined in the meta-analysis had positive mean effects on children's science achievement outcomes. The pooled average effect on science achievement was +0.33 SD. Using a calculation to convert effect sizes to more easily interpretable forms (Lipsey et al., 2012), this effect size translates to a scenario in which a typical student in the treatment group in an early childhood science educational intervention would be expected to rank at approximately the 63th percentile, as compared with a typical control group student who ranked at the 50^{th} percentile. Among studies that used randomized experimental designs, the pooled mean effect size on science achievement was +0.19 SD. Applying Kraft's (2020) benchmarks for effect size magnitudes in causal research on educational interventions, this would be considered a 'medium' effect.

The overall findings suggest that early childhood science interventions of the kinds evaluated were generally robust enough to induce a treatment contrast with typical science instruction at the grade levels we examined. Importantly, under typical conditions of the kinds likely experienced by the control groups, pre-K through first grade teachers receive little curricular support on providing science instruction in ways that are developmentally appropriate for young children and that are aligned with inquiry-centered pedagogical models (Davis et al., 2006; NASEM, 2022). In fact, Plumley (2019) found that 49% of elementary teachers frequently used science materials from internet sources such as Teachers Pay Teachers, which may or may not be high-quality, standards-aligned materials. Thus, teachers in typical control classrooms may have been expected to be using lower-quality instructional materials as compared to teachers receiving study treatments, contributing to a meaningful difference in treatment and control children's opportunities to learn science. The current findings, including the subset from randomized trials whose designs support causal inference, suggest that improving children's science learning outcomes in the early grades by intervening at the classroom level is possible.

At the same time, it is important to note that the types of early science educational interventions that tend to be evaluated in the literature using experimental and quasi-experimental methods are generally fairly intensive and deliberately aligned with research on how young children learn science. The programs evaluated in the meta-analysis offered teachers a median of 17 hours of PD, intended to expose children to a median of 25 hours of new curriculum, and focused on 3.2 of the five standards-aligned NASEM (2022) activities of investigation and design for young learners. By contrast, typical PD offered in schools tends to be of short duration, of variable quality, and may or may not be focused on research-based instructional practices (Hill et al., 2019; NASEM, 2022), As such, similar to other reviews (e.g., Garrett et al., 2019; Kraft et al., 2018) the findings imply that relatively intensive programs of the types that have undergone research evaluations tend to have positive effects, but these results do not necessarily generalize to other kinds of programs. We agree with others (e.g., Hill, 2004) that more research on commonly-implemented programs would enrich the literature.

Evidence for Moderators of Program Impacts on Science Achievement

We also sought to identify factors that predicted differences in programs' impacts. We initially hypothesized a number of potential moderators as predictors of program impacts, yet most of the variables we examined were not significantly related to effect size magnitudes. Rather, it appears that a wide variety of different kinds of programs can be effective at improving early childhood science learning outcomes, including those implemented in a range of settings and with and without each of the examined programmatic foci and teacher PD features.

The one examined programmatic moderator that was significantly related to effect size magnitudes for children's science achievement in the individual models was teacher PD contact hours above the median, and this relationship was negative. However, this association was not significant in the joint model. While a lack of a positive association between PD contact hours and student achievement outcomes may seem counterintuitive, given that more time spent on professional learning could reasonably be expected to lead to stronger teacher familiarity with an intervention and skill with its implementation (e.g., Desimone, 2009), several reviews have not found a clear association between PD duration and intervention impacts (e.g., Garrett et al., 2019; Lynch et al., 2019). One possibility is that PD duration may tend to be related to the complexity of the work that interventions require teachers to do (Garrett et al., 2019). Shorter PD programs may tend to focus on narrower instructional goals, and these programs may in turn be effective at helping students improve on more constrained, yet useful sets of outcomes (Garrett et al., 2019). At the same time, the median duration of the PD offered to treatment teachers in the interventions was 17 hours long, and many of the "below-median" offerings were still several hours. This is likely much longer than the norm for science PD offered to early grades teachers under typical conditions in schools. Thus, the pattern of findings we observed may not hold for the kinds of lower-intensity PD programs more commonly offered in schools.

An important concern from a STEM workforce perspective is improving science achievement in high-poverty settings, where students are at heightened risk for exiting the STEM pipeline and where opportunities to learn science have historically been constrained (e.g., NRC, 2007; NASEM, 2022). As one point of comparison, the magnitude of the science achievement gap between kindergarteners at roughly the 90th versus 10th percentiles of family income has been estimated at 1 SD (Curran, 2017). For the current meta-analysis, more than three-quarters of the studies that we identified in the literature were implemented in settings in which more than half of participating children were low-income. The moderator analysis results suggested that educational interventions to support early science learning did not differ significantly in their efficacy by sample poverty level. This finding is promising, suggesting that interventions can bolster early science achievement, including in settings serving predominantly low-income children who are important constituents of early childhood interventions. We did not observe significant differences in effect size magnitudes by urbanicity of the study setting nor participant grade level, however, most studies were conducted in urban settings. The evidence base on interventions conducted in rural, suburban, and mixed settings is relatively small. More research in these localities could inform our understanding of how local policy and resource environments outside of cities may influence programs' efficacy (e.g., Hill et al., 2023).

Effects on Teacher Outcomes and on Children's Literacy and SEL Skills

Although the number of studies that reported this information was relatively small, there is suggestive evidence of positive effects on students' literacy and SEL outcomes, and on teacher-level outcomes, including classroom instruction. Regarding teachers, the findings suggest that these interventions had positive pooled mean effects on teacher-level outcomes (+0.89 SD; k = 8). The pooled mean impact of early science interventions on measures of classroom instruction (+0.94 SD, k = 7) is larger in magnitude than the pooled mean impacts of science PD interventions found in other recent meta-analyses that included mostly studies conducted with older students (Kowalski et al., 2020, +0.34 SD; Lynch et al., 2025, +0.47 SD). One potential explanation for these findings is that early childhood teachers in control groups would typically be expected to have less exposure to training opportunities in science teaching methods during their preservice and in-service education as compared with their colleagues at the older grade levels (e.g., Barenthien et al., 2020) and thus new PD and curriculum interventions may introduce a particularly substantial treatment-control contrast for in-service teachers in the early grades. The number of studies that reported this information was small, hence caution is warranted in interpretation. However, the findings are consistent with a scenario in which early childhood educators' science instruction was malleable when teachers were given a variety of resources to help them improve, a promising finding that warrants future research.

There is also suggestive evidence about the potential for early science interventions to improve literacy outcomes alongside science learning, a possibility that is encouraging given that is has been of significant interest to the field (e.g., NRC, 2014). Among the twelve studies that examined literacy outcomes, the pooled mean impact estimate was 0.17 SD. Most, albeit not all, of the studies that measured literacy outcomes evaluated content literacy interventions.

Also encouragingly, the findings shown in Table 5 suggest that content literacy interventions also had positive mean impacts on science-related outcomes, in addition to the observed positive impacts on literacy. At the same time, the science-related skills targeted by content literacy interventions were often primarily, albeit not exclusively, a subset of skills that overlapped with literacy (e.g., science vocabulary, argumentative writing in science). Such skills are clearly quite important; at the same time, they encompass a partial subset of the full range of science-related skills that are posited in policy and standards documents to be valuable for young children to develop (e.g., NASEM, 2022; U.S. Department of Health & Human Services, 2024).

As such, early science content-literacy interventions of the kinds evaluated in the experimental and quasi-experimental research appear to be effective on average at improving the outcomes they are targeting, but children may benefit from participating in such interventions alongside opportunities to engage in other kinds of science learning, such as those that focus on SEPs, crosscutting concepts, and/or activities of investigation and design, in order to meet the full range of goals set out in early science learning standards.

While the number of studies that reported information on impacts on SEL outcomes was small, the pooled mean effect was positive (0.24 SD; k = 8), suggesting that early science programs can potentially foster SEL skills alongside science learning. SEL skills are of growing policy interest given the contemporary state of children's mental health (NASEM, 2023) and concerns about children's increasing difficulty regulating their behavior in ways necessary to maintain academic progress (e.g., Peetz, 2025). The included SEL outcomes ranged from measures of competence motivation and reading motivation to social skills. Given their importance, we urge future studies to report SEL impacts. Instructional approaches that target both science and SEL skills simultaneously also warrant experimentation (Garner et al., 2018).

Strengths, Limitations, and Future Opportunities for the Field

The current study findings indicate fruitful future opportunities for the field. First, like most meta-analyses in education research (e.g., Lynch et al., 2019; Scher & O'Reilly, 2009), study reports did not include all of the information we needed to conduct some analyses of interest, suggesting steps the field could take in the future to improve primary study reporting. In our moderator analyses, only one moderator was found to be significantly related to effect size magnitudes (i.e., PD duration below the sample median), and only in the individual model; however, measurement error could have contributed to this result. This is because information about some moderators of interest could only be weakly inferred from the study reports due to descriptions of intervention content that were lacking in detail (Lynch et al., 2025). Future studies would benefit from more consistent and detailed reporting of intervention content and experiences for both the treatment and control groups, in line with those recommended in recently-developed reporting standards for classroom interventions (Hill et al., 2025). Among these details we recommend that study authors report on the extent to which their intervention focused on crosscutting concepts; activities of investigation and design and/or SEPs (as appropriate for the sample); and disciplinary core ideas (NASEM, 2022; NGSS Lead States, 2013), as well as details of any teacher PD provided. With regard to any instructional coaching provided as a component of teacher PD, Kraft et al. (2018) recommend that study authors consistently report the coaching theory of action, dosage, and coach background and expertise.

There also appear to be interesting possibilities for more specific reporting in future studies on how school personnel interpreted and engaged with the new early grades science programs, both individually and collectively, and how school and district contexts may have influenced programs' implementation. Schools adapt to internal changes and external influences. Adding a new science program is much more than inserting a new module. There are potential ripple effects that range from individual staff members' uptake or resistance to the program, to alignment versus conflicts with other features of the school's program and culture. This complexity shapes the impact and durability of any program introduced to a school (Koh & Askell-Williams, 2021). For example, without more details in study reports, it is unclear how the approaches that school and program leadership took to programs' implementation may have interacted with schools' cultures and organizational climate among teachers.

We concur with others (Desimone & Pak, 2017; Hill et al., 2023; Hill et al., 2025; Lynch et al., 2019) in recommending that study authors report on contextual factors relevant to the study setting that could be expected to influence the intervention's effectiveness, such as resources provided for the intervention by schools and districts; teachers' background knowledge, experience, and attitudes; school climate; school and district leadership support for the intervention; and conflicting or competing initiatives. The field would also benefit more mixed-methods studies that include qualitative observations, surveys, and interviews, to unpack the leadership and school context mechanisms that affect implementation (e.g., Hill et al., 2023).

We also noted a paucity of studies in the literature on some topics of theoretical interest. Among the intervention foci with small presence in the data were the inclusion of family involvement; culturally relevant pedagogy or a social justice component; school partnership with an external community institution; and specific content areas, such as early engineering. It was not possible for us to model the effects of these features quantitatively due to their low prevalence. However, these would be potentially interesting topics for future syntheses to examine once a sufficient number of studies have been conducted that include these foci.

For example, only seven of the 31 experimental/quasi-experimental studies of classroom-based interventions that we identified reported including a parent/family involvement element in the program design. Because prior research and theory indicates that fostering strong mutual and reciprocal partnerships between home and school can benefit a variety of children's outcomes in the early grades (e.g., González et al., 2006; Krajcik et al., 2022), more research on early science educational interventions that involve both teachers and parents/caregivers would be valuable. Such work can focus researchers' attention on the science learning assets that children bring to school (e.g., Gunckel et al., 2024) and provide models that other researchers and practitioners could adapt.

A routine social policy fact is that the same type of intervention may be effective in some settings, where it is a good match for the needs and skills of participants and for the structure and resources of the local context, and less effective in others, where participant needs, the local context, and the intervention's structure or content are misaligned (Jacob et al., 2019). As such, we initially hoped to examine impacts disaggregated by specific student demographic groups (e.g., by baseline achievement, race/ethnicity, English language learners), but we did not find enough studies with outcomes disaggregated by demographic designations to permit such analyses. As argued by Hill et al. (2025), future studies could strengthen the evidence base by presenting effect size information disaggregated for specific groups relevant to the local setting.

We concur with Greenfield (2015) in suggesting that the field consider more widely adopting common assessments for early childhood science learning. We observed a marginally significant difference in outcomes by whether standardized versus researcher-developed measures were used; however, only eighteen effect sizes were derived from standardized assessments, while 57 were derived from researcher-developed assessments. The development and use of more common assessments has the potential to advance the field by defining shared understandings of desired outcomes of early science interventions, and clarifying how different kinds of programs are influencing outcomes on a common metric. Greenfield (2015) and colleagues have made important strides with the development and refinement of the Lens on Science tool. We also found that only a subset of studies reported information on interventions' impacts on classroom instruction. More comprehensive reporting on instructional outcomes could illuminate mechanisms by which early science educational interventions are achieving their effects. A generative stream of work has designed classroom observation tools for assessing early science instructional interactions (e.g., Greenfield et al., 2024; Kaderavek et al., 2015;

Vitiello et al., 2019). Standardized assessments could be administered alongside researcher-developed assessments, which provide valuable information about the specific outcomes targeted by individual interventions. Making information about existing early childhood science assessments more publicly accessible, such as via instrument repositories like EdInstruments, could also help early science researchers to more easily refine valuable outcome measures.

The promising findings also lead to questions about how early grades educators can build more time for high-quality science learning opportunities into children's days. School days are already quite full. However, policymakers could consider whether there are opportunities to find more time for science by reducing some time spent on activities that lack a strong evidence basis (i.e., efficiency gains). While some redundancy in school curricula is likely useful (Cohen-Vogel et al., 2021), Engel et al. (2013) found that teachers in U.S. kindergarten classrooms tended to spend excessive time re-teaching children early numeracy material that they had already mastered, and this repetition was detrimental to children's achievement. The authors concluded that "the vast majority of kindergarten students are exposed to large amounts of content that they do not benefit from in kindergarten" (Engel et al., 2013, p. 172). Meanwhile, some early grades classrooms may not include *any* regular instructional time for science (e.g., Anderson, 2012; Early et al., 2010; Piasta et al., 2014). In a national survey, about 40 percent of elementary teachers reported teaching science 3 or fewer days each week, and another ~40 percent reported teaching science some weeks but not every week (Banilower et al., 2018; Plumley, 2019). Reallocating even some of the time that children currently spend reviewing already-mastered content to science could represent a meaningful difference.

Some teacher PD time could also potentially be reallocated to early science instruction from common uses that lack strong research evidence, such as the isolated study of student test score data absent instructional supports (Hill, 2020). Furthermore, the increase in open-access high-quality curriculum materials available for early elementary grades (e.g., OpenSciEd) could broaden access to materials for teachers and students, and lower the barrier for time-pressed teachers who nonetheless hope to add more science learning opportunities to their classrooms.

In sum, we were able to systematically review and meta-analyze over two decades' worth of experimental and quasi-experimental research on the efficacy of early childhood science educational interventions. We argue that the current findings provide an existence proof for the notion that concentrated investments into bolstering curriculum and teacher professional development have the potential to help young children learn more science. The findings also point toward useful steps that the field can take to learn more from future research investments, with the ultimate goal of giving all young children a strong start in science that paves the way for future STEM opportunities and achievement.

References

Anderson, K. J. (2012). Science education and test-based accountability: Reviewing their relationship and exploring implications for future policy. *Science Education*, *96*(1), 104-129.

Banilower, E. R., Smith, P. S., Malzahn, K. A., Plumley, C. L., Gordon, E. M., & Hayes, M. L. (2018). Report of the 2018 NSSME+. *Horizon Research, Inc.*

Barenthien, J., Oppermann, E., Anders, Y., & Steffensky, M. (2020). Preschool teachers' learning opportunities in their initial teacher education and in-service professional development. *International Journal of Science Education*, *42*(5), 744-763.

Bustamante, A. S., Greenfield, D. B., & Nayfeld, I. (2018). Early childhood science and engineering: Engaging platforms for fostering domain-general learning skills. *Education Sciences*, *8*(3), 144.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, 54(4), 297–312. doi:10.1037/h0040950

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283-292.

Cohen-Vogel, L., Little, M., Jang, W., Burchinal, M., & Bratsch-Hines, M. (2021). A missed opportunity? Instructional content redundancy in pre-K and kindergarten. *AERA Open*, *7*.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Curran, F. C. (2017). Income-based disparities in early elementary school science achievement. *The Elementary School Journal*, *118*(2), 207-231.

Curran, F. C., & Kitchin, J. (2019). Early elementary science instruction: does more time on science or science topics/skills predict science achievement in the early grades?. *AERA Open, 5*(3).

Dabney, K. P., Chakraverty, D., & Tai, R. H. (2013). The association of family influence and initial interest in science. *Science Education*, *97*(3), 395-409.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181-199.

Desimone, L. M., & Pak, K. (2017). Instructional coaching as high-quality professional development. *Theory into Practice*, *56*(1), 3-12.

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

Early, D.M., Iruka, I.U., Ritchie, S., Barbarin, O.A., Winn, D.M.C., Crawford, G.M., Frome, P.M., Clifford, R.M., Burchinal, M., Howes, C., Bryant, D.M., and Pianta, R.C. (2010). How do pre-kindergarteners spend their time? Gender, ethnicity, and income as predictors of experiences in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, 25(2), 177–193.

Edwards, S., Skouteris, H., Cutter-Mackenzie, A., Rutherford, L., O'Conner, M., Mantilla, A., ... & Elliot, S. (2016). Young children learning about well-being and environmental education in the early years: A funds of knowledge approach. *Early Years*, *36*(1), 33-50.

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis) alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, *35*(2), 157-178.

Estrella, G., Au, J., Jaeggi, S. M., & Collins, P. (2018). Is inquiry science instruction effective for English language learners? A meta-analytic review. *AERA Open*, *4*(2).

Friedman-Krauss, A. H., Barnett, W. S., Garver, K. A., Hodges, K. S., Weisenfeld, G. G., Gardiner, B. A., & Jost, T. M. (2022). The state of preschool 2021: State preschool yearbook. National Institute for Early Education Research.

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.

Garner, P. W., Gabitova, N., Gupta, A., & Wood, T. (2018). Innovations in science education: Infusing social emotional principles into early STEM learning. *Cultural Studies of Science Education*, *13*(4), 889-903.

Garrett, R., Citkowicz, M., & Williams, R. (2019). How responsive is a teacher's classroom practice to intervention? A meta-analysis of randomized field studies. *Review of Research in Education*, *43*(1), 106-137.

Gibbons, L. K., & Cobb, P. (2017). Focusing on teacher learning opportunities to identify potentially productive coaching activities. *Journal of Teacher Education*, *68*(4), 411-425.

González, N., Moll, L. C., & Amanti, C. (2006). *Funds of knowledge: Theorizing practices in households, communities, and classrooms.* Routledge.

Gotwals, A. W., Wright, T. S., Pikus, A., Anderson, B. E., & Domke, L. (2025). Science talk from the start of school: The implementation of a formative observation tool to support professional learning. *The Elementary School Journal*, *125*(3), 000-000.

Greenfield, D. B. (2015). Assessment in early childhood science education. In *Research in early childhood science education* (pp. 353-380). Springer, Dordrecht.

Greenfield, D., Zan, B., McWayne, C., Harris, M., Alexander, A., Ochoa, W., & Mistry, J. (2024). Early childhood science practices observation tool (EC-SPOT): Assessing science practices across multiple classroom contexts. *International Journal of Science Education*, 1-19.

Gunckel, K. L., Davis, E. A., & Bautista, J. (2024). Moving beyond equity-as-access: Expanding what counts as science in the elementary classroom. *Science and Children*, *61*(1), 32-37.

Harrer, M., Cuijpers, P., Furukawa, T.A., & Ebert, D.D. (2021). *Doing meta-analysis with R: A hands-on guide*. Chapmann & Hall/CRC Press.

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.

Hill, C. J., Scher, L., Haimson, J., & Granito, K. (2023). *Conducting implementation research in impact studies of education interventions.* (NCEE 2023-005).

Hill, H. C. (2004). Professional development standards and practices in elementary school mathematics. *The Elementary School Journal*, 104, 215–231.

Hill, H. C. (2020, February 7). Does studying student data really raise test scores? *Education Week*.

Hill, H. C., Lynch, K., & Polanin, J. (2025). Structured reporting guidelines for classroom intervention research. *Journal of Research on Educational Effectiveness.*

Institute of Medicine and National Research Council. (2000). From neurons to neighborhoods: The science of early childhood development. The National Academies Press.

Jacob, R. T., Doolittle F., Kemple J., & Somers M.-A. (2019). A framework for learning from null results. *Educational Researcher*, 48(9), 580–589.

Judson, E. (2013). The relationship between time allocated for science in elementary schools and state accountability policies. *Science Education*, 97(4), 621–636.

Kaderavek, J. N., North, T., Rotshtein, R., Dao, H., Liber, N., Milewski, G., Molitor, S. G., &

Czerniak, C. M. (2015). SCIIENCE: The creation and pilot implementation of an NGSS-based instrument to evaluate early childhood science teaching. *Studies in Educational Evaluation*, 45, 27–36.

Kaderavek, J. N., Paprzycki, P., Czerniak, C. M., Hapgood, S., Mentzer, G., Molitor, S., & Mendenhall, R. (2020). Longitudinal impact of early childhood science instruction on 5th grade science achievement. *International Journal of Science Education*, *42*(7), 1124-1143.

Kennedy, M. M. (2016). How does professional development improve teaching?. *Review of Educational Research*, *86*(4), 945-980.

Koh, G. A., & Askell-Williams, H. (2021). Sustainable school-improvement in complex adaptive systems: A scoping review. *Review of Education, 9*(1), 281-314.

Kowalski, S. M., Taylor, J. A., Askinas, K. M., Wang, Q., Zhang, Q., Maddix, W. P., & Tipton, E. (2020). Examining factors contributing to variation in effect size estimates of teacher outcomes from studies of science teacher professional development. *Journal of Research on Educational Effectiveness*, *13*(3), 430-458.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547-588.

Krajcik, J. S., Miller, E. C., & Chen, I. C. (2022). Using project-based learning to leverage culturally relevant pedagogy for science sensemaking in urban elementary classrooms. In M. Atwater (Ed.), *International Handbook of Research on Multicultural Science Education* (pp. 913-932). Springer.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in US schools. *Educational Researcher*, 35(7), 3–12.

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

Lynch, K. (2024). The application of behavioral economics to teacher professional development. *Educational Researcher*, *53*(9), 508-520.

Lynch, K., Gonzalez, K., Hill, H., & Merritt, R. (2025). A meta-analysis of the experimental evidence linking mathematics and science professional development interventions to teacher knowledge, classroom instruction, and student achievement. *AERA Open*, *11*, 23328584251335302.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta–analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293.

National Academies of Sciences, Engineering, and Medicine. (2007). *Rising Above the Gathering Storm.* The National Academies Press. https://doi.org/10.17226/11463.

National Academies of Sciences, Engineering, and Medicine. (2022). *Science and engineering in preschool through elementary grades: The brilliance of children and the strengths of educators*. The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2023). *Addressing the long-term effects of the COVID-19 pandemic on children and families.* The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2025). *Equity in K-12 STEM* *education: Framing decisions for the future*. The National Academies Press.

National Research Council. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future.* National Academies Press.

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

National Research Council. (2014). *Literacy for science.* The National Academies Press.

O'Connor, G., Fragkiadaki, G., Fleer, M., & Rai, P. (2021). Early childhood science education from 0 to 6: A literature review. *Education Sciences*, *11*(4), 178.

Paprzycki, P., Tuttle, N., Czerniak, C. M., Molitor, S., Kadervaek, J., & Mendenhall, R. (2017). The impact of a framework-aligned science professional development program on literacy and mathematics achievement of K-3 students. *Journal of Research in Science Teaching*, *54*(9), 1174-1196.

Peetz, C. (2025). Is student behavior getting any better? *Education Week*.

Piasta, S.B., Pelatti, C.Y., and Miller, H.L. (2014). Mathematics and science learning opportunities in preschool classrooms. *Early Education and Development*, *25*(4), 445–468.

Plumley, C. L. (2019). 2018 NSSME+: Status of elementary school science. Horizon Research.

Pustejovsky, J., & Rodgers, M. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, 10(1), 57–71. https://doi.org/10.1002/jrsm.1332

Pustejovsky, J., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(3), 425-438.

Ravanis, K. (2017). Early childhood science education: State of the art and perspectives. *Journal of Baltic Science Education*, *16*(3), 284-288.

Raven, S., & Wenner, J. A. (2023). Science at the center: Meaningful science learning in a preschool classroom. *Journal of Research in Science Teaching*, *60*(3), 484-514.

Rodgers, M. A., & Pustejovsky, J. E. (2020). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, *26*(2), 141–160.

Savelsbergh, E. R., Prins, G. T., Rietbergen, C., Fechner, S., Vaessen, B. E., Draijer, J. M., & Bakker, A. (2016). Effects of innovative science and mathematics teaching on student attitudes and achievement. *Educational Research Review*, *19*, 158-172.

Scher, L., & O'Reilly, F. (2009). Professional development for K–12 math and science teachers: What do we really know?. *Journal of Research on Educational Effectiveness*, *2*(3), 209-249.

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

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37, 5–14.

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.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* National Academies Press.

StataCorp. (2021). *Stata 17 Base Reference Manual*. Stata Press.

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teachers good? A crosscase analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education*, *62*(4), 339-355.

Taylor, J. A., Kowalski, S. M., Polanin, J. R., Askinas, K., Stuhlsatz, M. A., Wilson, C. D., ... & Wilson, S. J. (2018). Investigating science education effect sizes: Implications for power analyses and programmatic decisions. *AERA Open*, *4*(3), 2332858418791991.

Trundle, K. C., & Saçkes, M. (Eds.). (2015). *Research in early childhood science education*. Springer.

U.S. Department of Health and Human Services. (2015). *Head Start early learning outcomes framework.* Administration for Children and Families, Office of Head Start.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. doi:10.18637/jss.v036.i03

Vitiello, V. E., Whittaker, J. V., Mulcahy, C., Kinzie, M. B., & Helferstay, L. (2019). Reliability and validity of the preschool science observation measure. *Early Education and Development*, *30*(2), 196-215.

What Works Clearinghouse. (2021). Reporting guide for study authors: Group design studies. Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. (2022). *What Works Clearinghouse standards handbook, version 5.0.* Institute of Education Sciences, U.S. Department of Education.

What Works Clearinghouse. (2014). Questions and answers from demystifying the What Works Clearinghouse: A webinar for developers and researchers.

Xie, Y., Fang, M., & Shauman, K. (2015). STEM education. *Annual Review of Sociology*, *41*, 331-357.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement. REL 2007-no. 033. Regional Educational Laboratory Southwest (NJ1).

You, H., Park, S., Hong, M., & Warren, A. (2024). Unveiling effectiveness: A meta-analysis of professional development programs in science education. *Journal of Research in Science Teaching*.

Table 1

Characteristics of Included Studies in the Meta-Analytic Sample

Characteristic	n	%
Study-level features		
Total	31	100.00
Type of publication		
Peer-reviewed journal article	22	70.97
Conference paper/poster	6	19.35
Technical report	2	6.45
Book chapter	1	3.23
Methodology		
Random assignment	24	77.42
Quasi-experimental	7	22.58
Evidence of differential attrition	2	6.95
Setting urbanicity		
Urban	18	58.06
Suburban, rural, or mixed	9	29.04
Not reported	4	12.90
Sample poverty level		
Majority low-income	24	77.42
Not majority low-income	7	22.58
Effect size-level features		
Total	173	100.00
Assessment type		
Standardized measures	92	53.18
Researcher-developed measures	81	46.82
Grade-level assessed		
Pre-kindergarten	93	53.76
Kindergarten	45	26.01
First grade	43	24.86

Note: Study-level features include two studies that include teacher-level impact estimates only.

Table 2

Characteristics of Included Interventions

Characteristic	N studies	% of studies or mean (SD)
Overall	31	100
Content area(s) addressed		
Science	31	100
Reading/ELA	13	41.94
Engineering	6	19.35
Mathematics	5	16.13
Domains of science-related content focus		
Life science	23	74.19
Earth and space science	13	41.94
Physical science	20	64.52
Engineering, technology, and the application of science	6	19.35
Programmatic foci of the interventions		
Number of focal activities of investigation and design	--	3.19 (1.78)
Focus on at least one crosscutting concept	19	61.29
Content literacy	10	32.26
Curriculum dosage hours	--	41.25 (70.72)
Features of the professional development		
Professional development contact hours	--	52.68 (109.42)
Focus on deepening aspects of teachers' content and/or pedagogical content knowledge	8	25.81
Teacher coaching and/or mentoring	13	41.94

Note: Studies include two studies that include teacher-level impact estimates only.

Table 3

Results of Fitting CHE Models Examining the Impacts of Early Science Classroom Interventions on Student Academic Outcomes

	All Student Outcomes		Achievement Outcomes		Science Achievement		Literacy Achievement	
Result	Robust Model	Controlling for design	Robust Model	Controlling for design	Robust Model	Controlling for design	Robust Model	Controlling for design
Intercept	0.254 ***	0.127 ***	0.274 ***	0.140 ***	0.331 ***	0.191 ***	0.165 +	0.143 +
Standard Error	(0.054)	(0.024)	(0.062)	(0.026)	(0.076)	(0.345)	(0.075)	(0.077)
Non-RCT		0.452 **		0.525 *		0.582 *		0.292 **
Standard Error		(0.112)		(0.171)		(0.240)		(0.077)
Variance								
Between study	0.063	0.004	0.089	0.017	0.110	0.032	0.056	0.054
Within study	0.037	0.038	0.038	0.039	0.071	0.073	0.009	0.009
Sample Size								
k studies	29		29		28		12	
n effect sizes	173		142		75		46	

Note: Robust standard errors are presented in parentheses.
Statistical significance is indicated by +p<.10, *p<.05, **p<.01, ***p<.001.

Table 4

Sensitivity of Pooled Mean Science Achievement Impact Estimates to Methodological, Publication, and Contextual Features

	Robust model			Controlling for design		
Characteristic	Point Estimate	Standard Error		Point Estimate	Standard Error	
Pooled effect among non-peer reviewed studies	0.167	(0.059)	*[a]	0.090	(0.062)	
Difference for peer-reviewed studies	0.218	(0.114)	+[b]	0.139	(0.082)	
Pooled effect among studies without differential attrition flag	0.341	(0.080)	***	0.196	(0.037)	***
Difference for studies with differential attrition flag	-0.198	(0.114)		-0.053	(0.054)	
Pooled effect among outcomes measured with standardized assessments	0.202	(0.054)	**	0.072	(0.064)	
Difference for outcomes measured with researcher-developed assessments	0.195	(0.091)	+	0.176	(0.087)	+
Pooled effect among outcomes for students in grades K or 1	0.341	(0.159)	+	0.156	(0.084)	+
Difference for outcomes for students in preschool/pre-K	-0.019	(0.241)		0.059	(0.148)	
Pooled effect among studies conducted in suburban, rural, or a mix of settings[c]	0.417	(0.199)	+	0.161	(0.095)	
Difference for studies conducted exclusively in urban settings	-0.086	(0.220)		0.053	(0.139)	
Pooled effect among studies not conducted with a majority low-income sample	0.256	(0.053)	**	0.178	(0.070)	*
Difference for studies conducted with a majority low-income sample	0.104	(0.115)		0.020	(0.093)	

Notes: Robust standard errors in parentheses.
[a] For each ‘pooled effect’ row, the statistical significance reflects the results of a hypothesis test that the point estimate is zero.
[b] For each ‘difference’ row, the statistical significance reflects the results of a hypothesis test that the difference between the magnitude of outcomes in the focal category versus the reference category is zero.
[c] Sensitivity analysis based on sample urbanicity excludes 4 studies (6 impact estimates) for whom this data was missing.
Statistical significance is indicated by +p<.10, *p<.05, **p<.01, ***p<.001.

Table 5

Results of Fitting a Series of CHE Models Investigating the Extent to which Programmatic Features Moderated Impacts on Science Achievement Outcomes

Characteristic	Bivariate moderation models (1)	(2)	(3)	(4)	(5)	(6)	(7)	Omnibus model
Intercept	0.306*	0.167+	0.164*	0.140+	0.283***	0.194***	0.207**	0.294
Standard error	(0.120)	(0.081)	(0.073)	(0.074)	(0.062)	(0.043)	(0.049)	(0.170)
Forms of activity of investigation and design	-0.034							-0.078
Standard error	(0.033)							(0.048)
Focus on at least one crosscutting concept		0.040						0.307
Standard error		(0.107)						(0.173)
Focus on developing content literacy			0.084					0.078
Standard error			(0.101)					(0.110)
Above median hours of curriculum exposure				0.094				0.081
Standard error				(0.122)				(0.138)
Above median hours of PD offered to teachers					-0.204*			-0.215
Standard error					(0.087)			(0.126)
PD focused on content knowledge and/or PCK						-0.001		0.055
Standard error						(0.064)		(0.142)
Teacher coaching and/or mentoring							-0.038	-0.053
Standard error							(0.094)	(0.080)

Note: The intercept represents the average impact estimate among non-RCT studies that do not include the moderating feature. Robust standard errors in parentheses. Statistical significance is indicated by +p<.10, *p<.05, **p<.01, ***p<.001. All models include a control for study design. Sample size for all models: 75 effect sizes from 28 studies.

Table 6

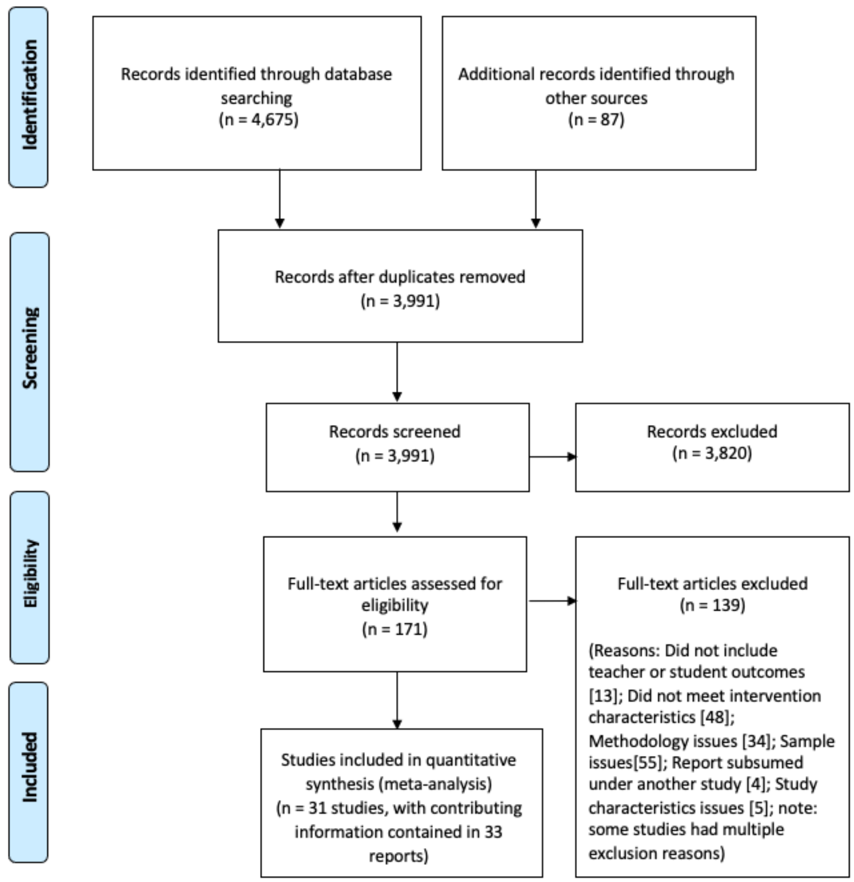
Results of Fitting CHE Models Examining the Impacts of Early Science Classroom Interventions on Teacher Outcomes and Student SEL Outcomes

		Teacher Outcomes		Student SEL Outcomes	
Characteristic		All Outcomes	Classroom Instruction Outcomes	Robust Model	Controlling for design
Intercept		0.887***	0.942***	0.237*	0.116
Standard error		(0.064)	(0.101)	(0.079)	(0.071)
Non-RCT					0.325+
Standard error					(0.107)
Sample Size					
	k studies	8	7	8	
	n effect sizes	44	30	28	

Notes: All teacher outcomes come from RCTs, so we are unable to control for research design in the teacher outcome models. Teacher outcomes in the 'all outcomes' column include measures of instruction, knowledge, and attitudes.
Statistical significance is indicated by +p<.10, *p<.05, **p<.01, ***p<.001.

Figure 1

PRISMA Screening Flowchart



Source: Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group (2009).
Note. PRISMA = Preferred Reporting Items for Systematic Reviews and Meta-Analyses.