Review of Educational Research
Month 202X, Vol. XX, No. X, pp. 1–46
DOI: 10.3102/00346543251389938
Article reuse guidelines: sagepub.com/journals-permissions

Exploring the Methodological Quality of Education Intervention Meta-Analyses: A Meta-Review

Marta Pellegrini
Department of Education, Psychology, Philosophy, University of Cagliari

Therese D. Pigott
Caroline Sutton Chubb
Natalie Pruitt
Hannah F. Scarbrough
College of Education and Human Development, Georgia State University

This meta-review explored the current practices used in education intervention meta-analyses in terms of systematic review procedures and meta-analysis methods. We reviewed 247 meta-analyses on the effects of K–12 school-based interventions on student academic achievement published after 2011. We found that many reviews were mostly consistent with several best practice recommendations for the review stage, including problem formulation, selection, and coding procedures. Reviews rarely preregistered their protocol or shared data, which reduces the transparency and reproducibility of the process. Best practice meta-analysis methods with robust consensus among methodologists were seldom used in our review sample. Recommendations for generating more credible and reproducible findings are provided. We also identify areas in need of more research and guidance, including how to conduct critical appraisal, how to deal with missing covariate data, and disciplined strategies to build meta-regression models.

Keywords: meta-analysis, methodological quality, transparency, reproducibility, achievement, program evaluation

Education research aims to influence policy and practice, providing quality evidence to improve student learning and developmental outcomes. Making decisions about education policy and practice requires reliable and valid primary studies and quality syntheses across these studies. Over the past two decades, the

methodology for systematic review and meta-analysis has rapidly evolved due to innovations in statistical methods for meta-analysis and improvements in tools for conducting stages of the systematic review, such as *Paperfetcher* (Pallath & Zhang, 2023) and *MetaReviewer* (Polanin et al., 2023). Tipton et al. (2019a, 2019b) provide an overview of the history of meta-analysis practice in psychology, education, and medicine, demonstrating extensive changes in meta-analysis modeling starting with the introduction of the use of methods for dependent effect sizes (Hedges et al., 2010). It remains unclear whether these new strategies have found their way to research synthesis in education.

Education researchers have increased interest in transparency and reproducibility in primary research (National Science Foundation & Institute of Education Sciences [NSF & IES], 2018). Similarly, systematic reviews and meta-analyses should adhere to principles of transparency and reproducibility (Lakens et al., 2016), particularly since most meta-analyses include public data that exists without privacy concerns. Though Polanin et al. (2020) found an increased use of open science practices in psychology meta-analyses, it is likely that education meta-analyses are also lacking in transparent and reproducible methods, limiting the utility and quality of meta-analyses for informing policy and practice.

This meta-review examines a sample of 247 education meta-analyses focused on K–12 interventions to increase academic achievement. We limit our review to meta-analyses on educational interventions because meta-analysis methods were first developed to examine intervention effectiveness (Glass, 1976), and much of the current guidance for conducting meta-analysis emphasizes the context of intervention syntheses (e.g., Higgins et al., 2019; Valentine et al., 2023). We examine the systematic review and meta-analysis methods and the transparency in methodological reporting. This meta-review also covers meta-analyses published across the education research journal landscape, including both journals that regularly publish meta-analyses and content-specific journals. Our goal is to understand how well education intervention meta-analyses focused on K–12 student achievement conform to current best practices and where more guidance is needed.

This meta-review aims to examine the following research questions:

1. To what extent are best-practice systematic review process methods used and reported in K–12 intervention meta-analyses?
2. To what extent are best-practice meta-analysis methods used and reported in K–12 intervention meta-analyses?
3. How do the systematic review and meta-analysis methods used in K–12 intervention meta-analyses differ across publication outlet, funding status, and time?

Contribution of the Meta-Review

As described in our protocol (Pellegrini et al., 2024), several meta-reviews have been published after 2010. In 2012, Ahn et al. (2012) conducted the most comprehensive meta-review to date, assessing the methodological quality of both

the *systematic review process* and *meta-analysis methods*. More recently, Nordström et al. (2023) investigated the risk of bias and open science practices (e.g., sharing data) used in education intervention meta-analyses published between 2019-2021. The meta-review included 88 reviews assessed in two stages using the shortened and full ROBIS checklists (Risk of Bias Assessment Tool for Systematic Reviews, Whiting et al., 2016). Eighteen reviews adhered to the shortened checklist assessment and were further assessed using the full ROBIS. Only a small portion of included reviews ($k = 10$) were judged to have a low risk of bias, six pre-registered a protocol, and three provided data. Polanin et al. (2017) conducted a meta-review of 20 education overviews of meta-analyses and determined that, while there are commonly agreed-upon standards for conducting and reporting systematic reviews, these guidelines were not consistently followed in overviews.

Our meta-review differs from previous meta-reviews in several ways. First, recent meta-reviews focused on a sample of reviews from specific journals or subjects (Hew et al., 2021; Nelson et al., 2022; Park et al., 2023), thereby limiting their scope. In contrast, our study offers a comprehensive and up-to-date meta-review of education intervention meta-analyses, similar to the work by Ahn et al. (2012). It examines reviews on the impact of K–12 interventions on student academic achievement, with no restriction to specific journals or subjects. Second, several studies have assessed specific stages of conducting systematic reviews, focusing either on review procedures (e.g., searching, selection) or on meta-analysis methods (e.g., effect size dependency, publication bias). Our meta-review assesses the quality of the systematic review process, meta-analysis methods, and open science practices.

Although checklists and assessments for the quality of systematic reviews exist (Shea et al., 2017; Whiting et al., 2016), most if not all of the current checklists were developed by health and medical researchers. Therefore, we developed our own codebook to reflect the specific context of education reviews. We began by coding items from existing quality assessment review tools and guidelines, codebooks used in previous education meta-reviews, and recent methodological research on meta-analysis practices (e.g., Pustejovsky & Tipton, 2022). See the preregistered protocol for a complete list of the references consulted during codebook construction (Pellegrini et al., 2024). Items were then grouped into categories based on the issue assessed and combined or revised as necessary. This process ensured a thorough evaluation of all procedures and methods essential to a high-quality review in education. We organized the framework of our review by three dimensions: (a) the quality of the *systematic review process*, (b) the quality of the *meta-analysis methods*; (c) the *significance for research use.* The systematic review process is related to the procedures and standards used to conduct the systematic review (e.g., search procedures) and open science practices (e.g., protocol preregistration, availability of data). The meta-analysis methods are related to procedures and standards on which the literature on meta-analysis appears to have a robust methodological consensus (e.g., handling effect size dependence, meta-regression to examine heterogeneity). Finally, the significance of the research use framework is related to the inclusion of analyses, results, and interpretations supporting the use of the research evidence in practice (e.g., under

which conditions the intervention works and implications for practice). These three framework dimensions reflect what we consider to be the current state of meta-analysis practice and its challenges. The stages of the systematic review process are well-established in the field and reflected in commonly used reporting guidelines such as PRISMA 2020 (Appelbaum et al., 2018; Page et al., 2021). Methods for searching, screening, and assessing the literature are also applied consistently across multiple types of systematic reviews. In contrast, meta-analysis methods have evolved significantly since well-known publications such as Lipsey and Wilson (2001) and Cooper (2017). We were interested in whether these innovations have influenced meta-analysis practice. The research-to-practice gap is well-established in the general education research, but little attention has been paid to the challenges of interpreting meta-analyses for practice. We were interested in documenting how published meta-analyses in the research facilitated the use of findings for practice. This review focuses on the quality of procedures, considering dimensions of the *systematic review process* and *meta-analysis methods*, while the significance of research use is addressed in Pellegrini et al. (2025).

Finally, this meta-review is timely, given recent international discussions about improving the credibility and relevance of review findings to better inform practice, policy, and decision-making (Maynard, 2024).

Expected Findings

In our preregistered protocol (Pellegrini et al., 2024), we hypothesized that recent meta-analyses published after 2020 would report a larger number of methodological characteristics and an increased prevalence of modern meta-analytical methods. We expected that meta-analyses would incorporate a larger number of best practices that have reached methodological consensus in social science applications of systematic reviews, such as methods addressing dependent effect sizes and multiple meta-regression to explore heterogeneity. Based on previous research (e.g., Tipton et al., 2019a) we expected fewer studies to consider more sophisticated methods which, despite gaining consensus among statisticians, have not been integrated into current practices. Among them, we expected few studies to use small-sample corrections, techniques to address multiplicity, and principled methods to handle missing data. We also anticipated that meta-analyses published in journals devoted to research syntheses (e.g., *Review of Educational Research, Educational Research Review*) would include more best practice methods than specialized-content journals. Journals focused on systematic review and meta-analysis typically recruit experts in systematic review and meta-analysis to serve on editorial boards and have more generous page limits for published manuscripts. Additionally, journals devoted to systematic review and meta-analysis may be more aware of current methodological standards and best practices. Finally, given the intense competition to secure funding, we expected that funded meta-analyses would have a higher quality than unfunded meta-analyses.

Methods

Our protocol was preregistered in the *Nordic Journal of Systematic Reviews in Education* (Pellegrini et al., 2024). The protocol and supplementary documents

are available on OSF at https://osf.io/2g7bk/. Deviations from the protocol are described in this section.

Inclusion Criteria

Our study included systematic reviews with meta-analysis only. Studies needed to report a summary effect size to be considered for inclusion. We only included meta-analyses synthesizing effects on student academic achievement. Other education-related outcomes, such as socio-emotional skills, attendance, dropout rates, computational thinking, and teacher outcomes, were excluded. We included reviews that evaluated the impact of school-based academic interventions. Thus, we excluded interventions that occur in school but are not directly related to academic learning, such as health interventions, after-school programs, physical activities, school structure, and social-emotional interventions. We only included motivation interventions when they focused on improving academic achievement. The eligible population consisted of K–12 students in general education. We only included pre-K or post-secondary when other eligible grade levels were also included. We excluded meta-analyses solely focused on special education populations and learning disabilities.

We restricted inclusion to reviews using group designs only (i.e., randomized controlled trials and quasi-experimental designs). We excluded correlational designs, single group pre-post designs, single-subject designs, and meta-analyses that combined different analysis designs (i.e., average effect size and model for heterogeneity) and were not conducted separately for studies with group designs. We excluded meta-analyses using single-subject experimental designs because the modeling strategies can differ significantly from those for group-design meta-analyses.

Considering the latest comprehensive review on the quality of education meta-analyses (Ahn et al., 2012), we included reviews published between January 2011 and September 2023 in English. Furthermore, Tipton et al. (2019b) indicated that 2010 was the beginning of a new phase of methodological growth in the field of meta-analysis. We restricted our search to papers published in peer-reviewed journals, excluding gray literature for two reasons: (a) Peer-reviewed studies have already passed an expert evaluation for their quality, and (b) some types of unpublished studies (e.g., conference papers) usually do not report the full methods section, and dissertations may follow guidelines and standards specific to their university. We assume that published meta-analyses are more likely to follow best practice guidance since many journals rely on expert methodologists as reviewers. The eligibility criteria are detailed in the protocol and summarized in Appendix A.

Search Strategy and Information Sources

We identified studies for the current meta-review using three search strategies. First, we searched the following electronic databases: Academic Search Ultimate, APA PsycInfo, Education Source, Education Resources Information Center (ERIC), Teacher Reference Center via EBSCOhost, Social Sciences Citation Index of Web of Science, and Science Direct. We identified search strings at the time of protocol preregistration (for complete search strings, see Appendix B),

including terms related to meta-analysis, intervention study, and participants, and adapted terms to fit database requirements as needed. Second, using *Paperfetcher* (Pallath & Zhang, 2023), we hand-searched the tables of contents in the following journals devoted to research syntheses or impact evaluations in education: *Review of Educational Research*, *Educational Research Review*, *Review of Research in Education, Campbell Systematic Reviews*, and *Journal of Research on Educational Effectiveness.* Third, we screened reference lists for all meta-analyses included in previous meta-reviews (see Pellegrini et al., 2024). The Nordström et al. (2023) review was identified after preregistration and, therefore, is not listed in our protocol.

Selection Process

A two-stage process was used for selecting relevant reviews. The title and abstract of the located meta-analyses were single-screened, only retaining meta-analyses in K–12 education for the full-text review. Reviews with titles and abstracts not explicitly stating meta-analysis or K–12 grade levels were retained for full-text review. The full texts of the retained studies were reviewed against the inclusion criteria by two authors independently. Disagreements were resolved via discussion among conflicted reviewers and an additional experienced reviewer. The initial inter-screener agreement was 81% with Cohen's $\kappa = 0.61$, which indicates substantial agreement (McHugh, 2012). After discussing the conflicts, 100% agreement was reached. Prior to screening, study selection guidelines were provided to the review team and continuously updated throughout the screening process. Training was conducted with review team members to practice screening on a weekly basis. The study selection process was conducted and documented in *Covidence* (https://www.covidence.org/). Guidelines for the selection process and the list of excluded studies with reasons for exclusion are accessible in the OSF repository.

Data Extraction Items

A draft codebook was included in the preregistered protocol and finalized during the initial stage of coder training. As described in the introduction, we developed a codebook draft based on guidelines, existing tools for review quality assessment, and codebooks of previous educational meta-reviews. We then combined codes for different tools that addressed the same element, ensuring retention of all relevant factors assessed by each tool. We organized the coded items in dimensions according to the characteristics assessed by: (a) *background information* (e.g., number of studies, journal); (b) *systematic review process*, which also includes open science practices; and (c) *meta-analysis methods*. The codebook is available in Appendix C.

In coding for the quality of the *systematic review process*, we extracted information related to each review stage. For problem formulation, we coded whether the included reviews addressed the two main goals of meta-analysis: measuring an average effect and exploring heterogeneity across interventions. We coded eligibility criteria based on the PICOS framework (Population, Intervention, Comparison, Outcomes, Study design; see, e.g., McKenzie et al., 2024) and we noted whether the authors excluded studies based on publication type. Publication

TABLE 1

Critical Appraisal Items and Definitions

Critical Appraisal Items	Definitions
Study design	Code designs as RCT or QED
Assignment level	Code cluster or individual-level random assignment
Publication status	Code studies as published or nonpublished
Baseline equivalence	Code whether the intervention and control groups were similar at the pretest
Attrition	Code whether the intervention and control groups had a similar number of students who dropped out of the study
Measurement	Code measurement characteristics, such as developer-made measures vs. independent measures
Control group type	Code whether the control group used BAU or an alternative intervention
Implementation fidelity	Code whether deviations from the intended intervention occurred
Interventionist	Code who delivered the intervention

bias or selective reporting bias exists in education research (Polanin et al., 2016). Searching for and including unpublished research decreases biased estimates of treatment effects.

All search practices used in the included reviews were coded, including how database searches were performed and what additional strategies were used, with particular attention to searching for gray literature. We coded the selection procedures used to conduct screening and full-text review and resolve conflicts, noting if any tools were used. We extracted similar information for the coding stage of each included review. Coding is an essential stage of meta-analysis as it is the process by which data is extracted to examine patterns or inconsistencies across studies. Coding manuals are preferred over narrative descriptions because they concisely display characteristics and their levels, along with explanations of what was coded and how.

To code critical appraisal practices, we first distinguished between a front-end and a back-end approach (Littell & Valentine, 2023). A front-end approach employs strict eligibility criteria, only including high-quality studies such as randomized controlled trials (RCTs), thus ensuring that all studies included can provide an unbiased estimate of the effect size. A back-end approach includes a wider range of study designs but then appraises each study with a set of items aimed at assessing the study's ability to provide an unbiased estimate of the target effect size. We also coded the use of a developed tool, such as Cochrane's checklists (e.g., risk of bias [RoB]), or author-selected methodological items (see Table 1). As suggested by Littell and Valentine (2023), critical appraisal should be tailored to the specific systematic review being conducted. One option is to use an existing tool, while another is to adapt one or create specific items to assess the extent to

which the included studies can provide trustworthy answers to the research questions.

Open science practices support the transparency and reproducibility of review procedures. We coded whether the authors preregistered a review protocol, shared data and statistical codes, and reported full search strings used in electronic databases. Key systematic review guidance, such as the APA Journal Article Reporting Guidelines (Appelbaum et al., 2018) and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, Page et al., 2021), note the importance of transparent reporting of the results of literature searching and screening results.

In coding for the quality of the *meta-analysis methods*, we focused on consensus procedures and standards supported by robust methodological consensus. Tipton et al. (2019a, 2019b) have described the following consensus points: use of all effect sizes in a model that accounts for effect size dependency, small sample corrections for meta-regression tests, multiple meta-regression (both continuous and categorical variables included in a single model) in modeling heterogeneity, distinction between confirmatory and exploratory analyses, and statistical adjustments for multiple comparisons (i.e., Bonferroni adjustments). Effect size dependency occurs when studies report multiple effect sizes and should be addressed using model-based methods. All strategies used by meta-analysts in the included reviews were coded. We noted the procedures for modeling heterogeneity using categories similar to Tipton et al. (2019a) to indicate subgroup analyses and ANOVA, historically used for categorical moderators, simple regression with a single continuous moderator, and multiple meta-regression with multiple moderators.

Other relevant coded features were measures of heterogeneity, statistical software used, and effect size calculations. We coded the type of software used, because traditional meta-analysis software may not include methods for modeling the multilevel and correlated structure of effect size data. Guidance about the use of effect sizes adjusted for baseline covariates is emerging (Hedges et al., 2023; Taylor et al., 2022), and we coded whether researchers were using these new procedures when coding adjusted effect sizes. In RCTs, baseline covariates increase the precision of the estimate and the power of the hypothesis test for the average treatment effect. In quasi-experimental designs (QEDs), the inclusion of covariates, especially the pretest of the outcome of interest, improves the internal validity by adjusting for any initial differences between groups (Hedges et al., 2023). We also examined how researchers used the critical appraisal results in their meta-analysis, for example, by excluding studies with a high risk of bias or accounting for their influence in the meta-analysis models.

Three common issues often encountered in meta-analyses were considered in our meta-reviews: publication bias, missing data on study characteristics, and outliers. We recorded whether our included meta-analyses mentioned those as concerns that may affect the findings, and what methods were used to address them.

Since our codebook was created using existing checklists and tools for assessing methods in systematic reviews and meta-analyses, it could also be applied to other educational topics beyond K–12 academic achievement. However, the following adjustments should be made when reviews are not designed to evaluate the

impact of interventions. Our codebook used the PICOS framework for coding inclusion criteria, which is only applicable to impact evaluations. Additionally, in coding critical appraisal, the available tools and the author-selected methodological items refer to characteristics commonly found in intervention studies using group designs.

Data Extraction Process

The procedure to code studies was organized according to the two dimensions of the codebook: the *systematic review process* and the *meta-analysis methods*. For the *systematic review process*, MP coded all studies and NP and HS double-coded 25% of the studies. For the *meta-analysis methods*, MP or TP coded all studies as first coder, and CC double-coded 25% of the studies. For both codebooks, the coder who was not involved in the conflicts reconciled them. The average inter-rater agreement for the review process form and the meta-analysis form was 85%–, with a range of 53% to 100%–, and 92%–, with a range of 60% to 100%, respectively (see complete list by item in the OSF repository). The lowest agreement (53%) was on an item recording the number of effect sizes included in the meta-analysis, an issue related to the reporting quality of the review and the number of analyses included. Other items with low agreement were also related to unclear reporting in the meta-analysis, such as details of the screening process used. Prior to data extraction, the review team pilot-tested the draft codebook on 10 studies and revised it as necessary (see Deviations From the Preregistration Protocol). This step helped to align coders and reach a high level of agreement. Regular meetings were held during data extraction to address coders' questions and enhance coding reliability. Data extraction was conducted in *MetaReviewer* (Polanin et al., 2023).

Some studies included multiple meta-analyses; for example, they calculated average effect sizes for different eligible outcomes. Since the methods used to conduct the systematic review and the meta-analysis were consistent across effect sizes, we coded each study once. For the number of included studies and the number of included effect sizes, we either extracted the most comprehensive data or the first reported data. For example, if a review reported the overall average effect size for academic achievement and separate averages for reading and mathematics, we extracted the academic achievement data. If only separate effect sizes by subject were available, we coded the numbers for the first analysis listed.

Data Analysis

We analyzed studies by providing descriptive statistics for coded characteristics. We calculated frequencies for categorical characteristics and means and standard deviations for continuous characteristics. We compared differences in methodological characteristics across time and journal type. Though planned, we did not explore the differences between funded and unfunded reviews because 47% of included reviews did not mention funding. We categorized studies based on the year of publication (2011–2015, 2016–2019, 2020–2023) and journal type (journals devoted to reviews vs. other journals). The journals categorized as devoted to reviews were hand-searched (i.e., *Review of Educational Research, Educational Research Review, Review of Research in Education, Campbell*

Systematic Reviews, Journal of Research on Educational Effectiveness). Based on our overall analysis, we identified key characteristics for which the included reviews did or did not meet high-quality standards. The systematic review process form examined the inclusion of gray literature, full reporting of search strings and PRISMA diagram, critical appraisal conduct, and open science practice. The meta-analysis form recorded the use of model-based methods for effect size dependency and effect size heterogeneity. Analyses and visualizations were performed using R Statistical Software (v. 4.4.2; R Core Team, 2024).

Deviations From the Preregistration Protocol

Deviations from the preregistered protocol occurred in the data extraction and data analysis stages. Adjustments were made to categories of data extracted to reduce the use of the "other" response option on the coding form. For example, the item regarding the approach to handling dependency was presented as a drop-down instead of a checkbox, which prevented the selection of multiple methods (e.g., averaged and shifting-unit-of-analysis). We coded these combinations under the "other" option and created categories a posteriori. We did not calculate the average quality score for the two assessed dimensions (*systematic review process* and *meta-analysis methods*) as stated in our protocol because we believe it is more beneficial to provide descriptive information about the practices and procedures used, along with an interpretation of the results. Additionally, creating an overall "quality score" by summing across a set of items would imply that each item is of equal importance to the construct of quality.

Results

We identified 2,003 potentially relevant records from database searches and 291 from hand-searches and retrospective reference harvesting. After removing 637 duplicates, the titles and abstracts of 1,657 records were screened. We retained and independently reviewed 486 records in full text. We excluded 239 reviews for several reasons, most commonly for ineligible outcome ($k = 84$), ineligible intervention ($k = 79$), ineligible student population ($k = 20$), and ineligible research design (such as a narrative review) ($k = 54$). A total of 247 reviews were ultimately included (see list of the included and excluded reviews in the OSF repository). Figure 1 shows a PRISMA flowchart documenting the search and selection process, and Table 2 reports background information from each of the included reviews. Across the included reviews, less than half were published in journals that were devoted to research syntheses (33%), such as *Review of Educational Research* or *Educational Research Review*, and less than half of the reviews reported receiving funding (43%). When looking at the year of publication, a greater proportion of our sample was published in more recent years, from 45 (18%) published between 2011–2015 to 126 (51%) between 2020–2023. The included reviews primarily evaluated intervention effects on reading (36%) or math (26%) outcomes or focused on academic achievement as a whole (34%). Not surprisingly, 60% of the reviews on technology interventions were published in the last four years (2020–2023). The reviews included an average of 46 studies, ranging from 5 to 406 studies, and had an average of 2.4 effect sizes.

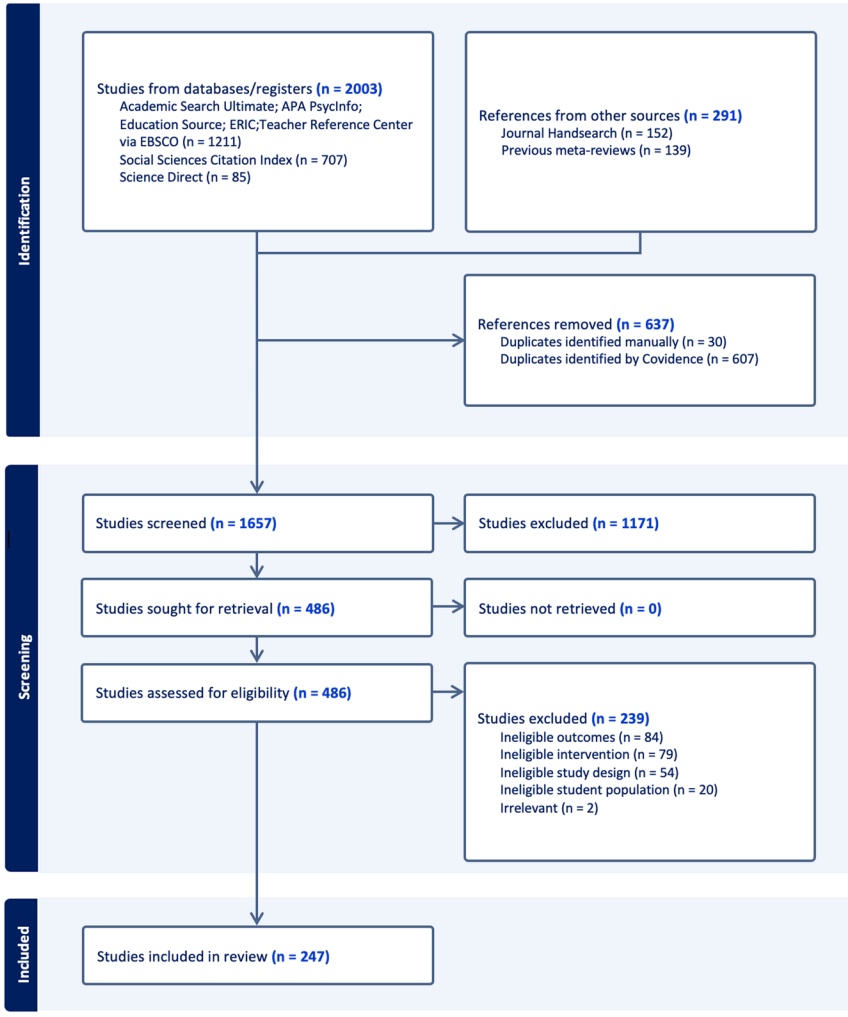

FIGURE 1. *Diagram of the selection process.*

We present our results by distinguishing between established practices with consistent adherence across reviews and practices that require improved attention from reviewers to enhance methodological quality. We first address the systematic review process dimension and its sections (i.e., problem formulation, inclusion criteria, searching, selection, coding, critical appraisal, and open science), followed by meta-analysis methods (i.e., synthesis, heterogeneity, and additional analyses). In each section, we discuss any results that differ across time periods and journal types. Otherwise, no differences were observed. The complete dataset is available in the OSF repository.

TABLE 2

Characteristics of the Included Meta-Analyses

Category	Number of Reviews ($k = 247$)
Number of included studies, *M* (SD)	46.3 (53.5)
Total number of effect sizes, *M* (SD)	130.8 (365.9)
Effect sizes per study, *M* (SD)	2.4 (2.4)
Journal type, *k* (%)	
Journals devoted to reviews	81 (33)
Other journals	166 (67)
Publication year, *k* (%)	
2011–2015	45 (18)
2016–2019	76 (31)
2020–2023	126 (51)
Funding, *k* (%)	
Funded	107 (43)
Not funded	25 (10)
Not mentioned	115 (47)

Systematic Review Process

Selected items about the systematic review process are provided in Table 3, with the full results provided in the OSF repository.

Research Questions Focus on Both the Average Effect Size and Heterogeneity

Most reviews addressed research questions related to the average effect and explored heterogeneity, indicating that both goals of a meta-analysis are widely shared by education researchers. The characteristics of the PICOS framework were included in the eligibility criteria in 86% to 99% of the reviews, except for the comparison group, which was present in only 56% of the reviews. Studies that did not mention the comparison group in the eligibility criteria placed no restrictions on the comparison condition, allowing both business-as-usual (BAU) and alternative practices to be eligible.

Multiple Search Strategies Are Common but Incompletely Reported

Included reviews used multiple methods for identifying relevant studies, such as database searches, prospective and retrospective searches, hand-searching journals, contacting authors, and search engines. Researchers used a median of two search strategies and a maximum of six. Unsurprisingly, the most frequently used search strategy was electronic database search (97%). Additionally, 60% of reviews retrospectively harvested references from included studies and prior meta-analyses. Rarer were strategies such as prospective forward citation searching (6%), contacting relevant researchers/authors (19%), and hand-searching journals (32%).

TABLE 3

Systematic Review Process

Systematic Review Process Item	*k* (%)
Problem Formulation	
Average effect research question	235 (95)
Heterogeneity research question	218 (88)
Inclusion Criteria	
Population	212 (86)
Intervention	245 (99)
Comparison	139 (56)
Outcome	227 (92)
Study design	235 (95)
Language	160 (65)
Timeframe	153 (62)
Publication type	97 (39)
Search Strategies	
Database search	239 (97)
Full search strings for each database[a]	22 (9)
Search engines	85 (34)
Hand search journals	79 (32)
Retrospective reference harvesting	148 (60)
Prospective forward citation search	14 (6)
Included and searched unpublished sources	149 (60)
Search dissertations[b]	131 (53)
Contact researchers[b]	46 (31)
Search websites of related associations[b]	25 (10)
Search independent research firms[b]	24 (10)
Conference proceedings[b]	9 (6)
Gray literature databases[b]	6 (2)
Selection Process	
Two-stage screening process	184 (75)
Report complete PRISMA flowchart	109 (44)
Independent double screening	51 (21)
Independent double full-text review	58 (24)
Reported tools used	20 (8)
Coding Process	
Narrative description of coded characteristics	186 (75)
Provide codebook/coding manuals	43 (17)
Independent double coding	122 (49)
Reported tools used	20 (8)

(continued)

Table 3. (continued)

Systematic Review Process Item	*k* (%)
Critical Appraisal	
Back-end approach	181(73)
Front-end approach	12(5)
Coded individual items	178(72)
Open Science Practices	
Preregistered protocol	9 (4)
Data available	15 (6)
Statistical code available	10 (4)

Note. [a] $k = 239$; [b] $k = 149$.

Very few (9%) reviews reported complete search strings for each searched database, though they were more commonly reported after 2016 (10%) compared to studies prior to 2016 (4%). Complete search strings were also reported more frequently in journals that specialize in systematic review (14%) than in content-related journals (7%). Most reviews (78%) only reported examples of search terms that may have been used in one or more databases. An additional 8% reported complete strings for some of the search databases, but the same percentage failed to report any search terms.

Searching and Inclusion of Gray Literature Varies

Searches for unpublished research were conducted in only 60% ($k = 149$) of the included reviews. Nearly 40% of the reviews excluded primary studies based on the publication type, either by selecting only peer-reviewed papers (33%) or excluding dissertations (6%). Of the reviews that included gray literature, the most common source of unpublished research was dissertations (131 of 149, or 88%). Other seldom-used strategies were contacting authors or prominent researchers (46 of 149, or 32%) or searching the websites of related associations (25 of 149, or 17%), independent research firms (24 of 149, or 16%), and conference proceedings (9 of 149, or 6%).

Selection and Coding Procedures Lack Transparency

Of the included meta-analyses, a little more than half (56%) reported a PRISMA diagram. However, only 44% of studies included a complete PRISMA flow diagram. Many PRISMA diagrams lacked elements such as exclusion reasons or specific gray literature sources or combined separate screening stages. The use of the PRISMA diagram increased steadily between 2011 and 2023, with only four studies out of 45 (9%) reporting a complete diagram between 2011 and 2015, 24 out of 75 (32%) in 2016–2019, and 81 out of 126 (64%) in 2020–2023. In the full-text review, an average of 58 (24%) of the studies used independent double-review, 19 (8%) used partial double-review, and a small percentage used single-review with or without validation. Similar selection procedures were used

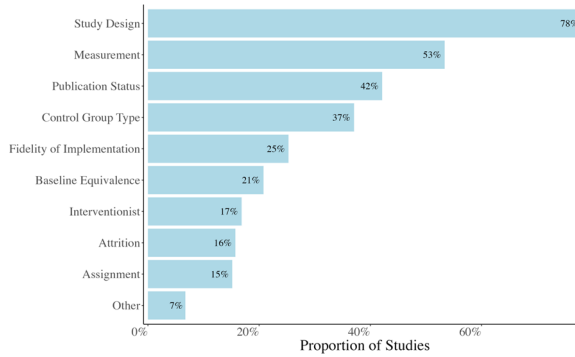

FIGURE 2. *Coded items for critical appraisal.*

during title and abstract screening. Of the 77 studies involving two independent reviewers in screening and/or full-text review—including partial double-review—59 of 77 (77%) reported accuracy in terms of inter-rater reliability or percentage of agreement, while only 26 of 77 (34%) described the process to reach consensus of disagreements. Additionally, a majority of the 247 included reviews (65%) did not report detailed procedures for full-text review.

Only 17% of the reviews provided a coding manual in tables or supplemental materials, while most reviews (75%) described the characteristics narratively. Procedures for coding were reported more thoroughly than those for study selection and screening. Most reviews involved partial or complete independent double-coding (71% or 175 reviews), of which 86% (150 out of 175) reported accuracy measures and 80% (140 out of 175) described the consensus process. Despite the increasing interest in machine learning and large language model tools that support study selection and the coding process, a very small percentage (8%) of reviews reported which screening or coding tools were used.

Meta-Analyses Typically Include Critical Appraisal of Included Studies

Most reviews (73%, 181 reviews) used the back-end approach (coding study quality elements), and 5% (12 reviews) used the front-end approach (including only high-quality studies). Among those that evaluated the methodological quality of primary studies ($k = 193$), the majority (92%, or 178 out of 193) selected specific items to code rather than using an existing tool. Figure 2 shows the characteristics most frequently coded. Other items coded by a small number of studies were sampling method, evaluator independence, and different effect size calculations. Researchers most frequently used the WWC guidelines and one or more of the Cochrane risk of bias tools (RoB1, RoB2, and ROBINS-I), with each of these tools selected by about 5% of the included 247 reviews.

Open Science Practices Are Rarely Implemented

We found that only nine of the 247 reviews (4%) preregistered a protocol. Of these, six were published in the journal *Campbell Systematic Reviews*, which

requires protocol submission before conducting the review. Eight of the nine studies preregistering a protocol were published in journals devoted to reviews after 2016. Also of note is the lack of transparency regarding conflicts of interest that might arise when review authors include their own research in a review. Only 2% of the included reviews stated a conflict of interest, while 33% specifically reported that there were none. Surprisingly, 65% of the reviews did not mention conflicts of interest at all, leading to a lack of transparency.

Of the reviews with published data, the complete dataset and statistical code were available for only 16 and 10 reviews, respectively, and mainly for reviews published after 2020. We observed no difference in the availability of data between systematic review journals and content-specific journals.

Meta-Analytic Methods

Selected items on the meta-analysis methods are provided in Table 4, with the full results provided in the OSF repository.

Meta-Analysts Use Diverse Methods for Computing Effect Sizes

At the study level, meta-analysts typically computed unadjusted effect sizes, ignoring baseline covariates (62%), while only 9% computed an effect size that accounted for at least the pretest of the outcome of interest. A substantial proportion of reviews (27%) combined pretest-adjusted effect sizes with unadjusted effect sizes. The choice of effect size types was typically based on the data available in the studies, with no distinct procedures reported for RCTs and QEDs.

Of the 35 reviews mentioning the inclusion of studies with cluster assignment, 24 used the adjustment suggested by Hedges (2007) and five used the adjustment reported in the *Cochrane Handbook* (Higgins et al., 2019). The methods discussed by Hedges (2007) focus on adjusting the standardized mean difference for cluster assignment using statistics typically reported in multilevel models (e.g., the ICC), while the *Cochrane Handbook* uses the design effect to correct effect-size standard errors.

Current Meta-Analysis Consensus Points Are Uncommon in Practice

Among the 88% ($k=217$) of reviews addressing multiple effect sizes within studies, 30% ($k=64$ of 217) used model-based methods accounting for dependency, which is the current best practice. Of those 64 studies, most ($k=36$) used robust variance estimation (RVE) with the correlated effects model and about one-quarter of the studies ($k=16$ of 64) employed the correlated and hierarchical effects model (CHE), with or without RVE (Pustejovsky & Tipton, 2022). As expected, we observed increased use of model-based methods over time, with 9% syntheses before 2015 using these methods, 25% between 2016–2019, and 17% between 2020–2023. We are unsure why the use of model-based methods decreased in 2020–2023. We speculate that during this time that coincided with the pandemic, many researchers turned to systematic reviews and meta-analysis work since they were unable to conduct primary research. These researchers may be less aware of innovations in meta-analysis modeling than those who used these methods prior to the pandemic.

TABLE 4

Meta-Analysis Methods

Meta-Analysis Methods Item	*k* (%)
Effect size calculation	
Pretest adjusted effect size	23 (9)
Unadjusted effect size	153 (62)
Adjusted and unadjusted effect size	67 (27)
Mentioned inclusion of studies with cluster assignment	35 (14)
Approach to handle dependency	
Dependency was handled in 88% (k = 217) of included reviews[a]	
Averaged or composite	53 (24)
Include all effect sizes ignoring dependency	21 (10)
Selected one	4 (2)
Subgroup approach/shifting unit of analysis	10 (5)
Model-based methods[a]	64 (30)
Multilevel meta-analysis	7 (11)
Multivariate meta-analysis	1 (2)
Multilevel meta-analysis with RVE	4 (6)
Correlated effects model (CE) with RVE	36 (56)
Correlated and hierarchical effects model (CHE)	2 (3)
Correlated and hierarchical effects model (CHE) with RVE	14 (22)
Heterogeneity	
Heterogeneity was assessed in 90% (k = 223) of included reviews[a]	
Q statistic reported	181 (81)
Tau-squared reported	74 (33)
Prediction interval reported	17 (8)
Publication Bias	
Publication bias was mentioned as a potential concern in 86% (k = 212) of included reviews[a]	
Funnel plot	150 (71)
Trim and fill methods	98 (46)
Fail-safe N (Rosenthal, Orwin)	93 (44)
Egger's regression	73 (34)
Indicator variable for published/unpublished	37 (18)
Selection models	6 (3)
Egger's sandwich	9 (4)
PET and PEESE	5 (2)
Outliers	
Outliers were mentioned as a potential concern in 48% (k = 119) of included reviews[a]	
Delete/winsorize outliers in main analysis	61 (51)
Sensitivity analysis without outliers	25 (21)
Sensitivity analysis with winsorized outliers	8 (7)

(continued)

Table 4. (continued)

Meta-Analysis Methods Item	*k* (%)
Leave-one-out analysis	16 (13)
Missing Data	
Missing data was mentioned as a potential concern in 21% (k = 51) of included reviews[a]	
Contact authors	12 (24)
List-wise deletion	23 (45)
Drop missing data moderators	9 (18)
Missing data as moderator level	6 (12)
Multiple imputation	4 (8)

[a]Percentages are based on the total number of studies of the category.

A range of strategies were employed when model-based methods for dependent effect sizes were not used. Of the 217 reviews that handled dependency, 24% ($k = 53$ of 217) averaged the effect sizes within studies, 10% ($k = 21$ of 217) treated all effect sizes as independent, 2% ($k = 4$ of 217) selected a single effect size, and 5% ($k = 10$ of 217) used a shifting-unit-of-analysis approach, conducting separate analyses on a single effect size per study. Another 24% ($k = 51$ of 217) used some combination of strategies, including shifting-unit-of-analysis, averaging, and adjusting the control group sample size when multiple treatments were included.

Most syntheses (80%) either used the software default settings for statistical tests or did not specify the settings used. Default settings typically rely on large-sample approximations, using normal or chi-squared reference distributions (Knapp & Hartung, 2003; Tipton, 2015). Small-sample adjustments for inferential procedures were implemented by 20% syntheses, most of which (88%) used RVE. Software for RVE, such as R *robumeta* and *clubSandwich* packages, usually implement those adjustments by default.

As expected, multiplicity was mentioned as a potential issue by only 6% of syntheses, with Benjamini-Hochberg and Bonferroni methods as the most used adjustments.

Heterogeneity Is Inconsistently Described Among Meta-Analyses

Nearly every meta-analysis (90%) examined heterogeneity, and many included multiple measures of effect size heterogeneity (see Table 4). Few reviews stated that they used a fixed effect model (4%), thus most K–12 intervention meta-analyses on academic achievement assume heterogeneity among studies. Most reviews reported the Q statistic (81%, or 181 of 223 reviews) and I^2 (63%, or 141 of 223). Despite the fact that most reviews used a random effects model, only one-third (33%, or 74 of 223) reported τ^2 (Borenstein et al., 2010). Only 8% (17 of 223) of studies reported a 95% prediction interval for the average effect size, a more meaningful description of heterogeneity (Borenstein, 2024).

In Exploring Heterogeneity, Meta-Regression With Multiple Covariates Is Rarely Used

Ninety-four percent of reviews (233 of 247) examined potential moderators of the effect size. Subgroup analysis was used to test categorical moderators, with 55% (128 of 233) using one-way ANOVA models and 21% (48 of 233) comparing average effect sizes between subgroups. Simple meta-regression, typically with a single continuous moderator, was used in 28% (66 of 233) of syntheses, whereas 26% (61 of 233) employed multiple meta-regression. Additionally, multiple meta-regression was more common in syntheses published after 2016 ($k = 54$ out of 205 or 26%) than between 2011–2015 ($k = 7$ out of 45 or 16%) and in syntheses published in journals devoted to reviews ($k = 29$ out of 81, or 36%) compared to content-related journals ($k = 32$ out of 166, or 19%). The small number of included studies may preclude the possibility of using multiple meta-regression. We then examined whether multiple meta-regression was consistently used in large reviews. We did not observe any clear pattern: some small reviews (e.g., fewer than 20 studies) used multiple meta-regression, while several large reviews with over 50 studies did not.

- Many included syntheses used multiple strategies to explore heterogeneity, such as a series of one-variable models and a single multiple meta-regression. When using meta-regression, we observed that the following approaches to model selection were the most used: sequential approaches in which models were compared with different sets of covariates and selected based on a specific criterion (e.g., forward and/or backward selection retaining only significant predictors);
- Forced-entry technique where all predictors were entered into the model at the same time. Sometimes, a small number of predictors were selected because of power issues;
- Separate multiple meta-regression models for groups of moderators (e.g., intervention, outcome), followed by a combined model retaining only significant predictors. In this case, the author often adjusted for potential methodological confounders;
- Analysis of single moderators followed by a model including all predictors to check the proportion of accounted variance and to control for confounders;
- Multiple meta-regression including only continuous variables;
- In some studies, the authors intended to use multiple meta-regression, but this was not possible because of issues with statistical power and multicollinearity.

Although it was not always clear how researchers built the models, the rationale for selecting predictors was typically based on theory. A few reviews (6%, or 13 of 233 reviews) distinguished between confirmatory and exploratory analysis. Since only a few reviews prospectively registered protocols, we cannot determine whether the authors followed planned analyses based on theoretical expectations or whether they engaged in a process to find statistically significant moderators.

A Range of Strategies Is Used to Examine How Critical Appraisal Items Relate to Heterogeneity

Of the 193 reviews that conducted critical appraisal, seven excluded the studies with critical risk of bias from the meta-analysis, while the majority of those 193 reviews (81% or 157 of 193) used quality ratings in the heterogeneity model. We observed three approaches that examine the relationship between study quality and effect size heterogeneity: (a) calculate and include a total quality score in the effect size model, (b) categorize the studies based on quality (e.g., low, medium, high) and include the categorical rating of study quality in the model, and (c) include key methodological characteristics as individual moderators in the effect size model.

Unclear Understanding of the Best Methods for Assessing Publication Bias

Our included reviews recognized the need to explore the potential impact of publication bias, with 86% (212 of 247) assessing selection reporting bias. On average, 2.4 methods were used, ranging from one to six methods per review. A combination of traditional methods to assess publication bias were most often reported, such as funnel plot (71%, 150 of 212), trim and fill methods (46%, 97 of 212), fail-safe N (44%, 93 of 212), Egger's regression (34%, 73 of 212), and using an indicator variable for unpublished vs. published as a moderator (18%, 37 of 212). More recently developed methods, such as selection modeling (3%, 6 of 212), PET/PEESE methods (2%, 5 of 212), and Egger's Sandwich (4%, 9 of 212), were also observed among our included reviews published after 2018.

Inconsistency Among Reviews for Handling Outliers and Missing Moderator Data

The included meta-analyses were split on attention to outliers, with 48% ($k = 119$) mentioning outliers as a potential issue. These findings may reflect the lack of guidance in the field on the importance of examining the reasons for outliers and addressing them. Half of the 119 reviews mentioning outliers (51% or 61 reviews) winsorized or deleted outliers in the main analysis. The remaining reviews conducted sensitivity analyses without outliers (21%, 25 of 119), performed leave-one-out analysis (13%, or 16 of 119) to check the sensitivity of results for every effect size, or included winsorized outliers (7%, 8 of 119).

In this meta-review, we focused on missing data for study characteristics and not effect sizes. Surprisingly, only 21% (51 of 247) of the reviews mentioned missing data as a potential issue at the moderator level. After attempting to find the missing information (24% (12 of 51 reviews), most reviews used list-wise deletion (45% 23 of 51), including only complete cases in effect size models. Other reviews dropped moderators with missing data (18%, 9 of 51) from any analysis or included a missing data indicator as a moderator alongside the moderator of interest (12%, 6 of 51). Since meta-analysts typically fit single-moderator models, effect sizes (cases) missing observations on that moderator were excluded, leading to varying numbers of effect sizes included in any one moderator analysis. This shifting-case-approach means that any single moderator analysis could include a different sample of effect sizes, limiting the ability to make inferences across all effect sizes in the data. Multiple imputation was performed

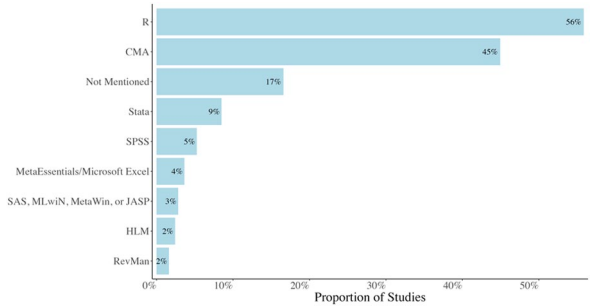

FIGURE 3. *Software used for statistical analysis.*

in only four reviews, and no reviews used other principled methods such as full information maximum likelihood (FIML). These results may also reflect the complexities of applying standard missing data methods to the case of meta-analysis.

The Choice of the Software May Limit the Meta-analysis Methods Used

In our included reviews, the most commonly used software was comprehensive meta-analysis (CMA) (45%), a point-and-click software that does not yet support multilevel models with a dependent effect size structure (see Figure 3). A small decrease in the use of CMA was recently registered, with 40% of reviews in 2020–2023 using CMA, 53% in 2016–2019, and 46% in 2011–2015. About 17% of studies mentioned the R program *metafor*, a commonly used and flexible package that includes functions for computing effect sizes and estimating complex effect size models. Fewer meta-analyses mentioned the R packages *robumeta* (11%) used for the correlated effects model or *clubSandwich* (6%) used for more complex multilevel effect size models. Less than 10% of reviews mentioned the use of STATA, SPSS, or other programs. A number of reviews (17%) did not mention the software used to conduct the analysis.

Discussion

This meta-review included 247 systematic reviews with meta-analysis focused on education interventions to improve K–12 academic achievement. Our meta-review examined a larger and more diverse set of education intervention meta-analyses compared to Ahn et al. (2012). Our research also covers a period of both rapid proliferation of systematic reviews and innovations in meta-analysis methods. The meta-analyses included in our review vary in their use of high-quality systematic review and meta-analysis methods. Some differences across years and journal types were found for specific items, such as reporting the PRISMA diagram and search strings and the use of model-based methods for dependent effect sizes, reflecting changing guidance in the field. Some methodological innovations we expected to see were uncommon, indicating that more effort is needed to ensure the dissemination of new methods. More detailed results are discussed later by dimension. Williams et al. (2022) and Vembye et al. (2024) are two

TABLE 5

Recommendations to Meta-Analysts

Key recommendations
Systematic review process
• Protocol preregistration
• Data and statistical code sharing
• Report complete search strings for database searches
• Use prospective search, retrospective search, contact authors, search engines, and websites of research centers and organizations
• Provide the codebook/coding manual with levels, explanations, and rationale for coded characteristics
• Use tools to support the review process (searching, selection, and coding) and report them clearly
Meta-analysis methods
• Use baseline covariate adjusted effect sizes (at least pretest) for QEDs and compromised RCTs. More generally, compute effect sizes that correspond as much as possible to the model used to make inferences in the primary analysis of the study
• Use appropriate adjustments for cluster studies
• Use model-based methods to handle dependent effect sizes
• Use multiple meta-regression (categorical and continuous moderators) to explore heterogeneity to control for confounders
• Use small-sample corrections for hypothesis tests
• Report τ^2 and the 95% prediction interval as measures of heterogeneity when using random-effects model
• Test and report the sensitivity of results to decisions made in the meta-analysis process
• Use scripting software for greater flexibility in the analysis and computational reproducibility. Use R as it is the most advanced software to handle dependent effect sizes

high-quality meta-analyses from our sample that incorporated most of the coded characteristics.

Table 5 reports key recommendations for generating more credible and reproducible findings. The recommendations are based on consensus points among methodologists that are still uncommon in practice, according to our results. Table 6 reports points for which more research and guidance are needed.

Systematic Review Stage

Our included meta-analyses follow many best practice recommendations for the review stage. Virtually all meta-analyses explicitly state research questions that include estimating the average effect size and describing the amount of heterogeneity among included primary studies (Tipton et al., 2023). Inclusion criteria are usually described by following a strategy such as the PICOS framework. Our sampled meta-analyses also conduct multiple strategies to search for eligible studies.

TABLE 6

Areas for Future Research and Guidance

Areas for more research and guidance
Systematic review process
• How to conduct a critical appraisal. This includes how to conduct it, when to use front-end or back-end approaches, the development of tools appropriate to primary studies in education, and what study characteristics should be coded
Meta-analysis methods
• Guidance for model building and covariate selection in multiple meta-regressions to explore heterogeneity
• How to treat the results of critical appraisal; when to exclude studies, when and how to select methodological characteristics to include in the model for heterogeneity
• What methods to use to assess selective reporting bias and how to choose among them
• What methods to use to handle missing covariate data and how to choose among them
• What methods to handle outliers when detected

Systematic reviews and meta-analyses should seek to avoid potential biases across the whole review process. The meta-analyses in our sample often failed to describe the process for screening eligible studies and coding included studies. Bias is also potentially introduced through incomplete searching and the failure to include the gray or unpublished literature. While dissertations are more easily searched through existing platforms such as ProQuest, few meta-analyses describe searching other sources of unpublished studies, such as research firm websites or conference proceedings.

We also found that transparency and reproducibility through open science practices are rare. A lack of transparency compromises the reproducibility of the procedures used, which in turn impacts the quality and comprehensiveness of the review results (Polanin et al., 2020). Few of the meta-analyses in our sample included a published or posted protocol. Less than half of our meta-analyses included a fully reported PRISMA diagram. Few meta-analyses provided the search terms used with each database and/or platform. Coding manuals were also absent in either the text or the supplemental materials. The lack of coding manuals prevents a deeper understanding of the moderators used in meta-analysis models and the sharing of information with others conducting similar synthesis projects. Additionally, few meta-analyses included the complete dataset, analysis plans, and/or meta-analysis codes to reproduce the results.

An area of confusion for many meta-analysts is selecting appropriate procedures for the critical appraisal of primary studies. While most meta-analyses used some type of appraisal for the quality of the included studies, a wide range of strategies are represented in our sample. The most widely known critical appraisal tools are those developed by Cochrane, including both versions of the risk-of-bias tools for randomized controlled trials (RoB1 and RoB2), and ROBINS-I for non-randomized studies. These tools were developed in the context of medical trials,

and many items, such as blinding, are irrelevant to education trials. Other meta-analyses coded a selected set of items, of which most focused on study design (RCT or QED) and publication status. Researchers also appear unclear about how to use the results of critical appraisal. The most common practices to control for study quality were including covariates in the heterogeneity model or excluding studies with poor quality.

Meta-Analysis Stage

We expected that more recent meta-analyses would acknowledge the presence of dependence among effect sizes nested within primary studies and the use of multivariate and multilevel models that more accurately reflect the structure of meta-analysis data. We observed that more recent meta-analyses tended to use modeling strategies that account for effect size dependence. However, the use of traditional meta-analysis methods (e.g., averaging effect sizes within studies) is still common in reviews conducted between 2020 and 2023.

Meta-analysts used diverse procedures for computing study effect sizes. Though several papers (Hedges et al., 2023; Taylor et al., 2022; What Works Clearinghouse [WWC], 2022) recently provided guidance for computing effect sizes adjusted for pretests and other baseline covariates, these practices are inconsistently used in meta-analyses. Covariate adjustments reduce bias in QEDs and improve precision in RCTs, although it is not strictly necessary in uncompromised RCTs. In addition, primary education studies frequently use ANCOVA or multiple regression to estimate the average effect of the intervention. Hedges et al. (2023) recommend that effect size computations closely align with statistical techniques used in the primary study analysis. For example, if a study uses ANCOVA with pretest as a covariate to account for its potential impact on the mean difference, the effect size should reflect this statistical adjustment.

Our included meta-analyses also used a range of methods for reporting and exploring heterogeneity. For example, few meta-analyses reported estimates of between-study variability (τ^2) despite most utilizing a random-effects model. A challenge for most meta-analysts includes strategies for fitting models of effect-size heterogeneity. Despite guidance on the use of meta-regression (Pigott & Polanin, 2020; Tipton et al., 2019a), many included meta-analyses used single-variable models. Disciplined methods for building effect size models are rare, providing another rationale for meta-analysts to provide *a priori* analysis plans and protocols.

Publication bias is recognized as an issue in meta-analysis. However, multiple methods are available to assess its presence, and confusion exists among meta-analysts on which methods are best to use. Furthermore, some methods to detect publication bias in reviews with dependent effect sizes were recently released (e.g., Chen & Pustejovsky, 2024; Pustejovsky, Citkowicz, & Joshi, 2025), and some time is needed for their dissemination. The knowledge regarding how to handle other issues that may affect the findings, such as missing covariate data and outliers, is less widespread.

The software used for meta-analysis can also limit or allow the use of correlated and hierarchical meta-analysis models and the computation of effect sizes

from more complex primary study designs. Point-and-click software, either designed explicitly for meta-analysis or more general programs such as SPSS, will limit the users' ability to fit more complex models. Scripting programs (e.g., R, STATA) provide greater flexibility for meta-analysts to fit appropriate advanced meta-analytic models. Among them, the R software stands out as the most comprehensive environment for meta-analysis, with packages such as *metafor* (Viechtbauer, 2010), *clubSandwich* (Pustejovsky, 2024), and *wildmeta* (Joshi et al., 2023) offering functionalities that are not fully available in other software. The Evidence Synthesis Hackathon (https://www.eshackathon.org) showcases R-based tools developed to support different stages of meta-analysis.

Limitations

We focused our meta-review on meta-analyses that examined the effectiveness of K–12 interventions for improving academic achievement. Thus, our findings apply to education intervention meta-analyses focused on this topic and cannot be generalized to meta-analyses conducted outside the school context and those examining relationships other than intervention effects. Other types of systematic reviews and meta-analyses exist in the literature, including those focused on estimating the correlation between two constructs, such as the association between academic achievement and socio-economic status or academic attitudes. These meta-analyses can include even more complexities, such as the inclusion of a wider range of study designs and questions about how to conduct a critical appraisal of primary studies. Our codebook could be adapted to examine such reviews.

Our eligibility criteria included only published meta-analyses, so our results do not apply to unpublished meta-analyses such as dissertations or other reports. We hypothesize that the quality of unpublished meta-analyses could vary widely. Not all education graduate schools offer a course on systematic review and meta-analysis, as evidenced by the interest in meta-analysis workshops offered by one of the study authors over the past decade. Adherence to best practices could be lower among dissertations that use meta-analysis or other types of unpublished reviews.

Other relevant characteristics need to be further explored, such as whether the reviews described the rationale used for coding specific study characteristics later tested as potential moderators. However, such information is ambiguous, and we believe it could not be extracted from our data with a high degree of reliability.

Recommendations

Due to advancements in systematic review and meta-analysis techniques, there is a clear understanding of best-practice methods and procedures in several review stages. Various guidelines are available to support the conduct of rigorous reviews and are continuously updated (e.g., Appelbaum et al., 2018; Page et al., 2021; Pigott & Polanin, 2020; WWC, 2022). Thus, the first general recommendation is to consult the guidelines and best practices before and during the process. We also recognize that meta-analysis methods are particularly complex and constantly evolving, which can result in some methods becoming outdated rather quickly.

This means that meta-analyses should be conducted by collaborative teams that include methodological experts in systematic reviews (such as information specialists), statisticians specializing in meta-analysis, and content-specific experts who can provide the necessary background and rationale to guide analysis and interpret results. A second aspect to consider is that disseminating new methods is crucial but remains a challenge in education research. In graduate school, emerging researchers have access to and the time to develop their methodological skillsets. It is possible that innovations in systematic review and meta-analysis methods over the past few decades have not found their way into graduate school training and have not been disseminated in ways that effectively reach meta-analysis researchers. It is unclear how best to ensure that new methods are shared widely across the field. We encourage researchers interested in meta-analysis methods to attend workshops offered by organizations focused on evidence synthesis (e.g., Campbell Collaboration) or by special-interest groups within societies (e.g., AERA-Systematic Reviews and Meta-Analysis SIG) and trainings conducted regularly by meta-analysis methodologists.

Systematic reviews and meta-analyses should adhere to principles of transparency and reproducibility to make results more relevant and credible for practice and policy decision-making. Meta-analysts should be as transparent as possible in reporting the process and results. Researchers should preregister a protocol, provide a coding manual for data extraction, and share data collected from studies and the statistical codes. This also includes publishing the list of studies identified with each search strategy and eligibility—in repositories or supplemental materials—along with the reasons for exclusion in the full-text review stage. Sharing statistical codes implies a shift from point-and-click software to scripting software (e.g., R software) that provides greater flexibility for fitting appropriate meta-regression models and supports computational reproducibility. We encourage researchers to adopt the described strategies as their standard practices.

Justifying all decisions made during the synthesis process is another way to enhance the credibility of findings. For example, single coding with validation from a second coder has recently emerged as a time- and resource-saving strategy compared to traditional double coding (Vembye et al., 2025), although more research is needed to assess the performance of validation under a range of conditions. Additionally, the use of machine learning tools to assist with searching and screening processes can be highly beneficial for reviewers in terms of saving time and resources (e.g., Ali et al., 2025; Zhang & Neitzel, 2024). However, their use should be clearly reported and well justified. Justifying decisions also applies to meta-analysis methods, showing whether the results are affected or unaffected by these choices. For example, meta-analysts in their review could have used different approaches to compute effect sizes (adjusted vs. unadjusted effect size) or to account for critical appraisal results—such as excluding studies with a high risk of bias, calculating a total quality score, or coding key methodological characteristics. These decisions should be supported

by the literature in the meta-analysis report and, when possible, the sensitivity of results to these decisions should be tested. As an example, Vembye et al. (2024) conducted a series of sensitivity analyses in which they repeatedly reestimated the average effect size while excluding certain categories of studies that may affect results (e.g., as nonrandomized studies or studies with a serious risk of bias) to assess the robustness of their findings.

Future Directions

Our findings suggest needed improvements in both meta-analysis guidance and methods. Two major improvements are needed. First, guidelines for systematic review and meta-analysis, such as the PRISMA and APA Journal Reporting Standards, provide clear recommendations for the review stage of a meta-analysis but are inconsistent with current modeling strategies. Recent improvements in meta-analysis methods need to be included in those guidelines, such as how to handle effect size dependency, report and describe effect size heterogeneity, and choose effect sizes that best reflect the analysis strategy used in the primary study (e.g., the use of effect sizes adjusted for baseline covariates). Second, we identify areas where current practices vary widely, highlighting the need for more research and/or guidance (see Table 6). For example, there is little direct guidance on how to address outliers and missing covariate data, on which methods should be preferred to assess publication bias, and on disciplined strategies for building meta-regression models. Recently, Pustejovsky, Zhang, & Tipton (2025) proposed a workflow for conducting preliminary analyses of meta-analytic databases, aimed at validating data integrity and guiding statistical modeling decisions. Critical appraisal is also a challenge for meta-analysts; a framework for thinking about study quality that better reflects the nature of educational studies is needed. Guidance will also be needed for how to apply newly developed programs and strategies that utilize large language modeling to automate or support aspects of systematic review. Recent scholarly work has explored programs assisting with citation searches, screening (e.g., Ali et al., 2025; Zhang & Neitzel, 2024), and coding (Wang & Luo, 2024).

Finally, as a third point, the quality of meta-analyses is strongly tied to the quality of the primary studies and how their methods and results are reported. Researchers conducting primary studies should place greater emphasis on accurate reporting of key study features, including research design, sample characteristics, context characteristics, statistical analyses, and data required for meta-analysis. By adopting these practices in high-quality primary studies, the evidence synthesized through meta-analyses would be even more relevant and generalizable, enhancing its practical value.

APPENDIX A

Summarized Inclusion Criteria

Inclusion	Exclusion
General populations of students in K–12. We include pre-K or post-secondary only when other K–12 grades are included.	Meta-analyses solely focused on special education populations (including learning disabilities).
School-based academic interventions, including motivation interventions, if they are focused on academic achievement.	Excluded interventions that may happen in school but are not directly related to learning an academic subject, such as health interventions, after-school programs, physical activities, school structure, social-emotional interventions.
Reports a summary effect size for student academic achievement.	Reports other education-related outcomes, such as socio-emotional skills, attendance, dropout rates, computational thinking, and teacher outcomes.
Studies using group designs (i.e., randomized controlled trials, quasi-experimental designs).	Correlational, single-group pre-post designs, single-subject design meta-analyses, as well as meta-analyses that combine different designs if the analyses (i.e., average effect size and model for heterogeneity) are not conducted separately for studies with group designs.
Published in English between January 2011 and September 2023.	Published before 2011.
Published in a peer-reviewed journal	Excludes unpublished studies (e.g., conference papers), dissertations, and reports.

Note: Adapted from the protocol (Pellegrini et al., 2024).

APPENDIX B

Search strings

Database	Search string	Limiters
Education Source Academic Search Ultimate APA PsycInfo ERIC Teacher Reference Center (via EBSCO)	(meta-analysis or meta-analytic or meta analysis) AND (K12 or K-12 or "elementary school" or "primary school" or "middle school" or "high school" or "secondary school" or kindergarten) AND (intervention or treatment or program or programme or experimental or experiment or RCT or trial or randomized)	Published 2011–2023 English language Academic journals and reports
Social Sciences Citation Index	((ALL=(meta-analysis or meta-analytic or meta analysis)) AND ALL=(K12 or K-12 or "elementary school" or "primary school" or "middle school" or "high school" or "secondary school" or kindergarten)) AND ALL=(intervention or treatment or program or programme or experimental or experiment or RCT or trial or randomized)	Published 2011–2023 English language Academic journals and reports
Science Direct	TITLE-ABSTR-KEY (meta-analysis or meta-analytic or meta analysis) AND ALL (intervention or treatment or program or programme or experimental or experiment or RCT or trial or randomized) AND (school or student or education)	Published 2011–2023 English language Subject area: social sciences Academic journals

Note: Search strings from the protocol (Pellegrini et al., 2024).

APPENDIX C

Codebook

Category	Item question	Options	Definitions and explanations
Background information			
Study information	Study ID from MR		
	Authors		
	Year of publication		
	Categories of publication year	2011–2015, 2016–2019, 2020–2023	
	Journal of publication		
	Type of the journal	Specialized, Other	Specialized = Journals devoted to reviews (i.e., *Campbell Systematic Reviews, Educational Research Review, Journal of Research on Educational Effectiveness, Review of Educational Research)* Other = all other journals
	Covidence ID		
	Was there funding received to support this study?	Yes, No, Not mentioned	
	If yes, report the name of the organization that funded the study		
	Is there a conflict of interest with this report?	Yes, No, Not mentioned	Yes = COI identified by the authors No = no COI identified by the coder Not mentioned = no COI mentioned in the study
	If yes, what is the nature of the COI?		Briefly describe it, e.g., meta-analysis on an intervention developed by the authors
	Type of interventions studied		Briefly describe the intervention studied, e.g., intelligent tutoring systems; elementary mathematics programs
	Categories of interventions studied		It groups interventions into categories
	Outcomes evaluated	Language arts, Reading, Mathematics, Science, Social studies, Academic achievement in general, Other (describe)	Select all that apply
	Number of studies included in the review		Total number of studies
	Number of effect sizes included in the review		The total number of effect sizes. When one effect size per study is used, we assume the same number as the studies

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
Systematic review process and open science practices			
Open science practices	Did the authors mention preregistration of the protocol?	Preregistered protocol in a peer-reviewed repository, Preregistered protocol in a non-peer-reviewed repository, No	Peer-reviewed repository=Protocol published in a peer-reviewed outlet (e.g., Campbell Collaboration or Cochrane Collaboration) Non-peer-reviewed=protocol available in an online repository (e.g., PROSPERO, OSF) No=no preregistration
	If the authors preregistered a protocol, did they mention any changes to the protocol?	Yes, No	
	Did the authors make the dataset available?	Yes in supplemental materials, Yes in an online repository, Yes on the authors' website, No, Other (describe)	Supplemental materials=published in journal Online repository=e.g., OSF Authors' website=including university storage or personal storage (e.g., Dropbox) No=no dataset shared or dataset available upon request to the authors
	Did the authors make the statistical code available?	Yes in supplemental materials, Yes in an online repository, Yes on the authors' website, No, Other (describe)	Supplemental materials=published in journal Online repository=e.g., OSF Authors' website=including university storage or personal storage (e.g., Dropbox) No=no dataset shared or dataset available upon request to the authors
Problem formulation	Did the authors provide an explicit research question/objective on estimating the average treatment effect?	Yes, No	
	Did the authors provide an explicit research question/objective on exploring the variation of the average effect across studies?	Yes, No	
Inclusion criteria	Did the authors list inclusion criteria?	Yes, No	Yes=criteria in any forms: list, text
	Did the authors list inclusion criteria for the population of interest?	Yes, No	Yes=grade level, demographics, country, etc.
	Did the authors list inclusion criteria for eligible interventions?	Yes, No	Yes=Type of interventions, definition of interventions
	Did the authors list inclusion criteria for the comparison or control condition?	Yes, No	Yes=description of the comparison, e.g., we compared tutoring with a nontutoring condition, business-as-usual, alternative program

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
	Did the authors list inclusion criteria for eligible outcomes?	Yes, No	Yes = description of dependent variable (e.g., mathematics achievement), types of measure, eligible assessment tools
	Did the authors list inclusion criteria for eligible study designs?	Yes, No	Yes = types of eligible research designs
	Did the authors list other inclusion criteria for eligibility of studies or reports?	Publication type, Language, Timeframe, No, Other (describe)	Publication type = any restriction, including only dissertations or only peer-reviewed journals
Searching	Which types of search strategies did the authors use?	Database search, Hand search journals, Retrospective reference harvesting, Prospective forward citation searching, Search engine, Other (describe)	Check all that apply. Retrospective reference harvesting = e.g., previous reviews, references of included studies. Prospective forward citation searching = studies that cite the included studies/previous reviews. Search engine = e.g., Google Scholar, Google
	Did the authors specifically search for gray literature?	Dissertations and/or theses, Contacting researchers, Searches of websites of independent research firms, Searches of websites of related associations, No, Other (describe)	Searches of websites of independent research firms = e.g., AIR, Rand, Mathematica Searches of websites of related associations = e.g., autism advocacy groups Under "other" (describe) = Conference programs, gray literature databases (e.g., Open Grey), strategies (e.g., funded project, program publishers, author name)
	Are search terms given for all database searches?	Complete strings for each database searched, Complete strings for some databases searched, Example/list of terms, Not reported	Complete strings = all strings used are reported in the text, supplemental materials, or online repository Examples/list of terms = just examples of words used, e.g., "tutoring" or "support" and "achievement"
Selection	Did the authors use a two-stage screening process?	Yes, No, Not mentioned	
	Which approach did the authors use for title and abstract screening?	Single-screening, Single-screening with validation from a second screener, Partial double-screening, Independent double-screening, Not mentioned	Single-screening = One screener Single-screening with validation from a second screener = One screener plus a second who validates decisions Partial double-screening = only a proportion of the studies were double-screened Independent double-screening = double-blinded screening of all records Not mentioned = no or unclear information

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
	Which approach did the authors use for full-text review?	Single-screening, Single-screening with validation from a second screener, Partial double-screening, Independent double-screening, Not mentioned	
	Did the authors report other procedures to check the accuracy of the selection stage?	IRR, Describe a consensus process, Not mentioned, Not applicable	Check all that apply: IRR = report an inter-rater reliability or agreement measure Describe a consensus process = e.g., reviewer training, meetings to review conflict and reach consensus Not mentioned = no information Not applicable = double-screening was not used
	Which tool did the authors use to conduct screening/full-text review?	Spreadsheet, Reference management software, Abstrackr, ASReview, Covidence, DistillerSR, EPPI-Reviewer, MetaReviewer, Rayyan, Not mentioned, Other (describe)	
	Did the authors report a PRISMA flowchart?	Yes, Incompletely reported, Not reported	Yes = only complete flowchart Incompletely reported = when the diagram does not show all stages of the process (initial number of located studies, retained number after screening and full-text review) or it does not follow PRISMA guidelines.
Coding	How did the authors report information about the characteristics coded?	Codebook, Narrative description of the characteristics, Not reported	Codebook = manual used to code studies with definitions. Codebook in the form of a table or spreadsheet may be reported in the text, supplemental materials, or online repository Narrative description of the characteristics = the description of items coded is embedded in the text No = not described
	Which approach did the authors use for coding?	Single-coding, Single-coding with validation from a second coder, Partial double-coding, Independent double-coding, Not mentioned	Single-coding = One screener Single-coding with validation from a second screener = One screener plus a second who validates decisions Partial double-coding = only a proportion of the studies were double-coded Independent double-coding = double-blinded coding for all records Not mentioned = no or unclear information

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
	Did the authors report other procedures to check the accuracy of coding?	IRR, Describe a consensus process, Not mentioned, Not applicable	Check all that apply: IRR = report an inter-rater reliability or agreement measure Describe a consensus process = e.g., reviewer training, meetings to review conflict and reach consensus Not mentioned = no information Not applicable = double-coding was not used
	Which tool did the authors use to conduct coding?	Spreadsheet, Reference management software, Abstrackr, ASReview, Covidence, DistillerSR, EPPI-Reviewer, MetaReviewer, Rayyan, Not mentioned, Other (describe)	
Critical appraisal	Did the authors mention taking into account the quality of primary studies, and if so, in which way?	Front-end approach, Back-end approach, No critical appraisal	Front-end approach = strict eligibility criteria to include only high-quality studies Back-end approach = broader eligibility criteria to include a larger number of studies that can be later assessed for quality No = no strict inclusion criteria + no critical appraisal (Littell & Valentine, 2023).
	Did the authors use a critical appraisal tool of primary studies?	No, Cochrane RoB 1.0, Cochrane RoB 2.0, Conn's (2017) Quality Index, Cook's et al. (2015) WI for group design, Effective Public Health Practice Project, JBI institute, Newcastle-Ottawa Scale, ROBINS-I, WWC standards, Medical Education Research Study Quality Instrument (MERSQI), Cook's et al. (2015) QI for group design, Selected items by authors, Other (describe)	Check all that apply. They may use more than one if they include both RCTs and non-RCTs Selected items by authors = methodological characteristics coded. Reviews using a front-end approach may also assess critical appraisal.

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
	If "selected items by authors," which characteristics did they code?	Study design, Assignment level, Publication status, Baseline equivalence, Attrition, Measurement, Control group type, Implementation quality, Interventionists, Other (describe)	Check all that apply Study design = design type, e.g., RCT or QED Assignment level = cluster or individual level random assignment Publication status = studies as published or nonpublished Baseline equivalence = whether the intervention and control groups were similar at the pretest Attrition = whether the intervention and control groups had a similar number of students who dropped out of the study Measurement = measurement characteristics, e.g., developer-made measures vs. independent measures Control group type = whether control group used BAU or an alternative intervention Implementation fidelity = whether deviations from the intended intervention occurred Interventionist = who delivered the intervention
	If the authors used a critical appraisal tool, did they describe the overall results of the critical appraisal either in a table or in narrative form?	Yes, No, Not applicable	Yes = descriptive results of the overall quality or description of methodological characteristics. We did not consider here whether they included those characteristics in the moderator analysis Not applicable = no critical appraisal.
Meta-analysis methods			
Synthesis methods	Average effect size of the main meta-analysis		Consider the average effect size for academic achievement. In case of more than one average effect size, consider the first one. If they reported FEM and REM, take REM results.
	Which measure of uncertainty did the authors provide for the average effect size? Standard error of average effect size of the main meta-analysis (if provided) Confidence interval of average effect size of the main meta-analysis (if provided)	Standard error, Confidence interval, Not reported	Check all that apply.
	How did the authors calculate study effect sizes?	Unadjusted effect size, Adjusted effect size, Not mentioned	Check all that apply. Adjusted ES = effect sizes are adjusted for pretest differences Unadjusted ES = pretest differences not accounted for. Posttest effect size only

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
	Did the authors mention they included studies with clusters as the unit of assignment?	Yes, No	
	If yes, did the authors adjust effect sizes and variances for clustering?	Yes mentioned Cochrane Handbook, Yes mentioned Hedges (2007), No	
	Did the authors say they used fixed-effects model or random-effects model?	Unweighted mean, FEM, REM, Unclear/Not reported	Unweighted mean = simple average of effect sizes FEM = Fixed-effects model REM = Random-effects model Unclear/Not reported = when we could not understand what was used. When different models were used, we selected REM and coded the other model under "Additional analyses"
	Did the authors provide a justification for the model used?	Yes, No	This is related to the previous question.
	Did the authors mention there were dependent effect sizes within studies?	Yes, No	Yes = explicit mention of multiple effect sizes per study No = no mention. Authors may not mention multiple effect sizes but specify a strategy (e.g., averaging effect sizes within studies). In this case, we selected "no," but we coded the strategy used in the next item.
	How did the authors handle dependent effect sizes?	Averaged/composite, Selected one, Subgroup approaches/ shifting unit-of-analysis approach, Model-based methods, Include all/multiple treatments effect sizes ignoring dependency, Adjust the sample size in the control group, Other, Combination of two or three methods (describe), Not mentioned	Averaged/composite = averaging effect sizes within each study. Selected one = choosing one effect size among the ones reported in the study. Subgroup approaches/shifting unit-of-analysis approach = creating subgroups of effects based on a characteristic (e.g., outcome) to perform several meta-analyses Model-based methods = including all effect sizes of studies. Include all/multiple treatments' effect sizes ignoring dependency = including and performing the meta-analysis with all effect sizes ignoring the dependency of the effect sizes within the studies Adjust the sample size in the control group = split the "shared" control group into two or more groups with smaller sample size and include two or more (reasonably independent) comparisons (old version of Cochrane Handbook) Combination of two or three methods = specify the methods

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
	If using a model-based method for dependent effect sizes, which one did the authors use?	Multivariate meta-analysis, Multilevel meta-analysis, Correlated Effects model with RVE, Hierarchical Effects model with RVE, Correlated and Hierarchical Effects (CHE) model, CHE with RVE	Multivariate meta-analysis = Effect sizes within studies are correlated Multilevel meta-analysis = three-level structure of effect sizes nested in subgroups nested in studies Correlated effects model with RVE = same assumption of multivariate meta-analysis with the use of robust variation estimation hierarchical effects model with RVE = same assumption of multilevel meta-analysis with the use of RVE Correlated and hierarchical effects model = multivariate and multilevel error structure. Assume same correlation between effect sizes across studies Correlated and hierarchical effects model with RVE = multivariate and multilevel error structure with RVE. Assume same correlation between effect sizes across studies (Hedges et al., 2010; Pustejovsky & Tipton, 2022; Vembye et al., 2023)
	Which statistical test did the authors use for the mean effect size?	t-test, Permutation test, Default, Other (describe)	Default = Default is implied when no adjustment or option is reported.
	Did the authors use a small sample correction for the test of the significance of the average effect size?	Yes, No, Not mentioned, Other (describe)	Yes = the statistical test reported or the default option of the software packages implies a small-sample correction No = the statistical test reported or the default option of the software packages is a large-sample test assuming a normal sampling distribution Not mentioned = when the statistical test is not reported
	Did the authors use a method for multiple comparisons correction?	Ad hoc methods, Benjamini-Hochberg, Bonferroni, Permutation test, No, Other (describe)	No = Not report or unclear
	Which statistical software did the authors use?	Comprehensive Meta-Analysis, R_robumeta, R_metafor, R_clubSandwich, R_meta, R_dmetar, R_metaSEM, HLM, STATA, SAS, SPSS, RevMan, Not mentioned, Other (describe)	
Heterogeneity	Did the authors test for heterogeneity of the mean effect size?	Yes, No	

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
	If yes, which measure of heterogeneity did the authors provide?	Q, I-squared, tau-squared, 95% Prediction interval, Not reported, Other (describe)	Q=(Cochran's Q) weighted sum of squared differences between individual study effects and the pooled effect across studies, with the weights being those used in the pooling method I-squared=proportion of total variation in effect sizes due to heterogeneity rather than chance Tau-squared=between-study variance in random-effects meta-analysis models Prediction interval=the range within which we expect the true effect of a future study to lie, accounting for both within- and between-study variability
	If there is heterogeneity, how did the authors model it?	Subgroup one-way, Subgroup multiway, ANOVA one-way, ANOVA multiway, Meta-regression simple, Meta-regression multiple, Not modeled, Not reported, Other (describe)	Subgroup one-way=compare categories of a moderator (one at a time) reporting the average effect sizes for each category with no statistical test Subgroup multiway=compare categories of moderators (multiple moderators at a time) reporting the average effect sizes for each category with no statistical test ANOVA one-way=compare effect sizes of categories of a moderator (one at a time) using statistical test ANOVA multiway=compare effect sizes of categories of moderators (multiple moderators at a time) using any statistical test Meta-regression simple=one moderator at a time Meta-regression multiple=multiple moderators at a time Not reported=no information on how they modeled heterogeneity Not modeled=no moderator analysis Tipton et al. (2019a)
	If the authors used multiple meta-regression, how did they describe the process of model building?		Quotes from the paper. Report information for model building (only include studies that used meta-regression).
	Did the authors distinguish between confirmatory and exploratory analysis?	Yes, No	Confirmatory=with a priori hypotheses Exploratory=based on data and interpreted with appropriate caution Tipton et al. (2019a)

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
Additional analysis	Did the authors mention publication bias as a potential issue?	Yes, No	
	If the authors assessed publication bias, which methods did they use?	Any fail-safe *N*, Funnel plot, Egger's Regression Test, Egger Sandwich, Selection modeling, Unpublished vs. published as a moderator, Begg and Mazumdar's rank correlation test / Kendall's Tau rank test, Precision Effect Test (PET) and Precision Effect Estimation With Standard Error (PEESE), Trim and fill method, Not mentioned, Other (describe)	Fail-safe *N*=estimates how many additional studies with null effect would be needed to bring the overall meta-analysis result to non-significance Funnel plot=a scatterplot of effect sizes versus their standard errors. Asymmetry may indicate bias Egger's regression test=a statistical test based on a funnel plot that checks for asymmetry by regressing the effect sizes on their standard errors Egger's sandwich=a variant of Egger's test that adds sandwich-type variance estimators to account for effect size dependencies, providing a more robust estimate Selection modeling=models how publication bias may have affected the observed data by estimating a selection mechanism, often based on the statistical significance of studies Unpublished vs. published as a moderator=publication status used in the moderator analysis. Begg and Mazumdar's Rank Correlation Test / Kendall's Tau Rank Test: non-parametric tests for publication bias Precision Effect Test (PET): Tests whether there is an association between study effect sizes and their precision (inverse of standard error), specifically looking for evidence of bias in studies with lower precision. Precision Effect Estimation with Standard Error (PEESE): An extension of PET that estimates the effect size while adjusting for publication bias by using the inverse of the squared standard error Trim and fill method: A method that "trims" asymmetric studies from the funnel plot, then "fills" in symmetrical missing studies.
	Did the authors mention missing data as a potential issue?	Yes, No	

(continued)

APPENDIX C (continued)

Category	Item question	Options	Definitions and explanations
	If the authors handled missing moderator data, which procedures did they use?	Contact the authors, Check other published syntheses/ reports, Ad hoc methods, Include "missing" as level of moderators, Drop moderators with too much missing data, After compare results for different models, Multiple imputation, Other principled methods, Other (describe)	Check all that apply Do not consider when they contacted authors for missing data to calculate effect sizes, but only for moderators. Ad hoc methods = list-wise deletion/complete case analysis Multiple imputation = replacing each missing value with a set of plausible values to create multiple complete datasets Other principled methods = e.g., full information maximum likelihood (FIML)—instead of discarding incomplete cases, FIML directly calculates the likelihood of the observed data, accounting for missingness under the assumption that data are missing at random
	Did the authors mention outliers as a potential issue?	Yes, No	
	If yes, how did the authors detect and handle outliers in the analysis?	Delete/winsorized outliers in the main analysis, Sensitivity analysis without outliers, Sensitivity analysis with winsorized values, Leave-one-out analysis, Sensitivity analysis with outliers (main analysis deleting outliers), Other (describe)	Delete/winsorized outliers in the main analysis = removing or replacing values by the nearest values within a specified percentile range in the main meta-analysis Sensitivity analysis without outliers = performing an additional analysis without outliers to check for the sensitivity of results Sensitivity analysis with winsorized values = performing an additional analysis with winsorized values to check for the sensitivity of results Leave-one-out analysis = sensitivity analysis technique where each study/effect size is systematically removed from the analysis one at a time Other = e.g., studies that checked for outliers but did not detect any
	Did the authors check the sensitivity of results to critical appraisal?	Study quality ratings used in the model for heterogeneity, Studies with critical risk of bias deleted from the model, Did not include study quality ratings in models of heterogeneity, Did not estimate models of heterogeneity, No critical appraisal conducted in the study	Check all apply. Study quality ratings used in the model for heterogeneity = overall quality rating or categorical moderators coded (e.g., randomized vs. non-randomized; level of attrition) included in the model for heterogeneity Studies with critical risk of bias deleted from the model = delete those studies before conducting the analysis Did not include study quality ratings in the models of heterogeneity = perform the moderator analysis without including quality appraisal results Did not estimate models of heterogeneity = no moderator analysis performed No critical appraisal conducted in the study
	Did the authors perform any other additional analysis? Describe.		For example, FEM

Declaration of Conflicting Interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Drs. Pigott and Pellegrini are authors of three reviews included in our meta-review, and as such, they were not involved in the screening and coding of these three studies.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the Young Researchers Mobility Program 2023 (Programma Mobilità Giovani Ricercatori [MGR]) of the University of Cagliari.

ORCID iDs

Marta Pellegrini https://orcid.org/0000-0002-9806-3231
Therese D. Pigott https://orcid.org/0000-0002-5976-246X
Caroline Sutton Chubb https://orcid.org/0009-0005-9542-8492
Natalie Pruitt https://orcid.org/0009-0001-0358-7420
Hannah F. Scarbrough https://orcid.org/0009-0003-7806-1049

References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, *82*(4), 436–476. https://doi.org/10.3102/0034654312458162

Ali, F., Tan, A. S.-C., & Wang, S. J.-W. (2025). Can machine learning help accelerate article screening for systematic reviews? Yes, when article separability in embedding space is high. *Research Synthesis Methods*, *16*(1), 194–210. https://doi.org/10.1017/rsm.2024.16

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, *73*, 3–25. https://doi.org/10.1037/amp0000191

Borenstein, M. (2024). Avoiding common mistakes in meta-analysis: Understanding the distinct roles of Q, I-squared, tau-squared, and the prediction interval in reporting heterogeneity. *Research Synthesis Methods*, *15*(2), 354–368. https://doi.org/10.1002/jrsm.1678

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Chen, M., & Pustejovsky, J. E. (2024, October 25). *Adapting methods for correcting selective reporting bias in meta-analysis of dependent effect sizes*. https://doi.org/10.31222/osf.io/jq52s

Conn, K. M. (2017). Identifying effective education interventions in sub-Saharan Africa: A meta-analysis of impact evaluations. *Review of educational research*, *87*(5), 863–898. https://doi.org/10.3102/0034654317712025

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2015). CEC's standards for classifying the evidence base of

practices in special education. *Remedial and Special Education*, *36*, 220–234. https://doi.org/10.1177/0741932514557271

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach*. Sage Publications.

Glass, G. V. (1976). Primary, secondary, and meta-analysis. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.2307/1174772

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

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

Hedges, L. V., Tipton, E., Zejnullahi, R., & Diaz, K. G. (2023). Effect sizes in ANCOVA and difference-in-differences designs. *British Journal of Mathematical and Statistical Psychology*, *76*(2), 259–282. https://doi.org/10.1111/bmsp.12296

Hew, K. F., Bai, S., Dawson, P., & Lo, C. K. (2021). Meta-analyses of flipped classroom studies: A review of methodology. *Educational Research Review*, *33*, 100393. https://doi.org/10.1016/j.edurev.2021.100393

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane Handbook for systematic reviews of interventions*. John Wiley & Sons.

Joshi, M., Pustejovsky, J., & Cappelli, P. (2023). *Wildmeta: Cluster wild bootstrapping for meta-analysis*. R package (Version 0.3.2). https://CRAN.R-project.org/package=wildmeta

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*, 24. https://doi.org/10.1186/s40359-016-0126-3

Littell, J. H., & Valentine, J. C. (2023). Unit 5: Data extraction and coding. In J. C. Valentine, J. H. Littell, & S. Young (Eds.), *Systematic reviews and meta-analysis: A Campbell Collaboration online course*. Open Learning Initiative. https://oli.cmu.edu/courses/systematic-reviews-and-meta-analysis/

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. SAGE Publications, Inc.

Maynard, R. (2024). Improving the usefulness and use of meta-analysis to inform policy and practice. *Evaluation Review*, *48*(3), 515–543. https://doi.org/10.1177/0193841X241229885

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282. https://doi.org/10.11613/bm.2012.031

McKenzie, J. E., Brennan, S. E., Ryan, R. E., Thomson, H. J., Johnston, R. V., & Thomas, J. (2024). Chapter 3: Defining the criteria for including studies and how they will be grouped for the synthesis. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, & V. A. Welch (Eds.), *Cochrane Handbook for systematic reviews of interventions* (Version 6.5). Cochrane. www.training.cochrane.org/handbook

National Science Foundation & Institute of Education Sciences (NSF & IES). (2018). *Companion guidelines on replication & reproducibility in education research*. https://www.nsf.gov/pubs/2019/nsf19022/nsf19022.pdf

Nelson, G., Park, S., Brafford, T., Heller, N. A., Crawford, A. R., & Drake, K. R. (2022). Reporting quality in math meta-analyses for students with or at risk of disabilities. *Exceptional Children*, *88*(2), 125–144. https://doi.org/10.1177/00144029211050851

Nordström, T., Kalmendal, A., & Batinovic, L. (2023). Risk of bias and open science practices in systematic reviews of educational effectiveness: A meta-review. *Review of Education*, *11*(3), e3443. https://doi.org/10.1002/rev3.3443

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 1–11. https://doi.org/10.1136/bmj.n71

Pallath, A., & Zhang, Q. (2023). Paperfetcher: A tool to automate handsearching and citation searching for systematic reviews. *Research Synthesis Methods*, *14*(2), 323–335. https://doi.org/10.1002/jrsm.1604

Park, S., Lee, Y. R., Nelson, G., & Tipton, E. (2023). Four best practices for meta-analysis: A systematic review of methodological rigor in mathematics interventions for students with or at risk of disabilities. *Learning Disability Quarterly*, *47*(4), 234–246. https://doi.org/10.1177/073194872311851

Pellegrini, M., Day, E., Scarbrough, H.F., Pigott, T. (2025). A meta-review of education meta-analyses: Relevance, applicability, and accessibility of findings. *AERA Open*, Advance online publication. https://doi.org/10.1177/23328584251389562

Pellegrini, M., Pigott, T., Chubb, C. S., Day, E., Pruitt, N., & Scarbrough, H. F. (2024). Protocol for a meta-review on education meta-analyses: Exploring methodological quality and potential significance for research use in practice. *Nordic Journal of Systematic Reviews in Education*, *2*, 76–102. https://doi.org/10.23865/njsre.v2.6169

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46. https://doi.org/10.3102/0034654319877153

Polanin, J. R., Austin, M., Peko-Spicer, S., Ebersole, C., Michaelson, L., Clements, J. Soule, C., Lee, S., Ezzat, Y., Williams, S., Mitchell, S., & Williams, R. T. (2023). *MetaReviewer* (Version 1.2) [Computer software]. American Institutes for Research. https://www.metareviewer.org/

Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Perspectives on Psychological Science*, *15*(4), 1026–1041. https://doi.org/10.1177/1745691620906416

Polanin, J. R., Maynard, B. R., & Dell, N. A. (2017). Overviews in education research: A systematic review and analysis. *Review of Educational Research*, *87*(1), 172–203. https://doi.org/10.3102/0034654316631117

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research*, *86*(1), 207–236. https://doi.org/10.3102/0034654315582067

Pustejovsky, J. E. (2024). *clubSandwich: Cluster-robust (Sandwich) variance estimators with small-sample corrections*. R package (Version 0.5.11). https://CRAN.R-project.org/package=clubSandwich

Pustejovsky, J. E., Citkowicz, M., & Joshi, M. (2025, May 28). *Estimation and inference for step-function selection models in meta-analysis with dependent effects*. https://doi.org/10.31222/osf.io/qg5x6_v1

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

Pustejovsky, J. E., Zhang, J., & Tipton, E. (2025, May 7). *A preliminary data analysis workflow for meta-analysis of dependent effect sizes*. https://doi.org/10.31222/osf.io/vfsqx_v1

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing (Version 4.4.2) [Computer software], Vienna, Austria. URL https://www.R-project.org/

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ, *358*, j4008. https://doi.org/10.1136/bmj.j4008

Taylor, J. A., Pigott, T., & Williams, R. (2022). Promoting knowledge accumulation about intervention effects: Exploring strategies for standardizing statistical approaches and effect size reporting. *Educational Researcher*, *51*(1), 72–80. https://doi.org/10.3102/0013189X211051319

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

Tipton, E., Bryan, C., Murray, J., McDaniel, M. A., Schneider, B., & Yeager, D. S. (2023). Why meta-analyses of growth mindset and other interventions should follow best practices for examining heterogeneity: Commentary on Macnamara and Burgoyne (2023) and Burnette et al. (2023). *Psychological Bulletin*, *149*(3–4), 229–241. https://doi.org/10.1037/bul0000384

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019a). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*(2), 180–194. https://doi.org/10.1002/jrsm.1339

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019b). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*(2), 161–179. https://doi.org/10.1002/jrsm.1338

Valentine, J. C., Littell, J. H., & Young, S. (Eds.). (2023). *Systematic reviews and meta-analysis: A Campbell Collaboration online course*. Open Learning Initiative. https://oli.cmu.edu/courses/systematic-reviews-and-meta-analysis

Vembye, M. H., Christensen, J., Mølgaard, A. B., & Schytt, F. L. W. (2025). Generative pre-trained transformer models can function as highly reliable second screeners of titles and abstracts in systematic reviews: A proof of concept and common guidelines. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000769

Vembye, M. H., Pustejovsky, J. E., & Pigott, T. D. (2023). Power approximations for overall average effects in meta-analysis with dependent effect sizes. *Journal of Educational and Behavioral Statistics*, *48*(1), 70–102. https://doi.org/10769986221127379

Vembye, M. H., Weiss, F., & Hamilton Bhat, B. (2024). The effects of co-teaching and related collaborative models of instruction on student achievement: A systematic

review and meta-analysis. *Review of Educational Research*, *94*(3), 376–422. https://doi.org/10.3102/00346543231186588

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.

Wang, X., & Luo, G. (2024, May 2). *MetaMate: Large language model to the rescue of automated data extraction for educational systematic reviews and meta-analyses*. https://doi.org/10.35542/osf.io/wn3cd

What Works Clearinghouse (WWC). (2022). What Works Clearinghouse Procedures and Standards Handbook, Version 5.0. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). *This report is available on the What Works Clearinghouse website at* https://ies.ed.gov/ncee/wwc/Handbooks

Whiting, P., Savović, J., Higgins, J. P., Caldwell, D. M., Reeves, B. C., Shea, B., Davies, P., Kleijnen, J., Churchill, R., & ROBIS group. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, *69*, 225–234. https://doi.org/10.1016/j.jclinepi.2015.06.005

Williams, R., Citkowicz, M., Miller, D. I., Lindsay, J., & Walters, K. (2022). Heterogeneity in mathematics intervention effects: Evidence from a meta-analysis of 191 randomized experiments. *Journal of Research on Educational Effectiveness*, *15*(3), 584–634. https://doi.org/10.1080/19345747.2021.2009072

Zhang, Q., & Neitzel, A. (2024). Choosing the right tool for the job: Screening tools for systematic reviews in education. *Journal of Research on Educational Effectiveness*, *17*(3), 513–539. https://doi.org/10.1080/19345747.2023.2209079

Authors

MARTA PELLEGRINI is an associate professor at the Department of Education, Psychology, Philosophy, University of Cagliari, Via Is Mirrionis 1, Cagliari 09127, Italy; email: marta.pellegrini@unica.it. Her research focuses on methods for evidence-based education, including systematic reviews with meta-analysis and impact evaluations on educational interventions. She is also a visiting professor at the Center for Research and Reform in Education, Johns Hopkins University.

THERESE D. PIGOTT is a distinguished university professor in the College of Education and Human Development, Georgia State University, 30 Pryor St. SW, Suite 300 Atlanta, GA 30303; email: tpigott@gsu.edu. Her research focuses on methodological advances in meta-analysis, including methods for missing data, statistical power, and outcome reporting bias. She is also an adjunct professor at the Knowledge Centre for Education, University of Stavanger.

CAROLINE SUTTON CHUBB is a PhD candidate in education policy studies and a graduate research assistant in the Educational Policy Studies Department in the College of Education and Human Development, Georgia State University, 30 Pryor St. SW, Suite 450, Atlanta, GA 30303; email: cchubb1@student.gsu.edu. Her research focuses on meta-analysis methods, best practices, and policy implications

NATALIE PRUITT is a doctoral student in research, measurement, and statistics in the Educational Policy Department in the College of Education and Human Development at Georgia State University, 30 Pryor St. SW, Suite 450, Atlanta, GA 30303; email:

npruitt4@student.gsu.edu. Her research focuses on meta-analysis methods and risk-of-bias assessment.

HANNAH F. SCARBROUGH is a PhD candidate in education policy studies and a graduate research assistant in the Educational Policy Studies Department in the College of Education and Human Development, Georgia State University, 30 Pryor St. SW, Suite 450 Atlanta, GA 30303; email: hscarbrough1@student.gsu.edu. Her research focuses on research synthesis methods and the use and communication of research evidence for practice and policymaking.