Nordic Journal of Systematic Reviews in Education
Volume 2 | 2024 | pp. 76–102

Protocol

Protocol for a Meta-Review on Education Meta-Analyses: Exploring Methodological Quality and Potential Significance for Research Use in Practice

Marta Pellegrini[1]*, Terri Pigott[2], Caroline Sutton Chubb[2], Elizabeth Day[3], Natalie Pruitt[2] & Hannah F. Scarbrough[2]
[1]*University of Cagliari, Italy;* [2]*Georgia State University, United States;*
[3]*University of Oregon, United States*

ABSTRACT
This protocol presents the planned methods and procedures for a meta-review that explores the methodological quality of education meta-analyses and their potential significance for research use. An ever-increasing number of meta-analyses are published every year while at the same time new approaches to enhance the validity and reproducibility of systematic reviews are in constant development. We aim to conduct a meta-review to examine the current practices of education meta-analyses in terms of review procedures, meta-analysis methods, and strategies for making meta-analyses useful for non-research audiences. We will focus on meta-analyses including randomized controlled trials and quasi-experimental designs on the effects of K-12 school-based academic interventions on student academic achievement. A comprehensive search will be conducted to retrieve all studies that potentially meet our criteria. Study features will be coded to evaluate three main dimensions: the quality of the review process, which includes the practices used at each systematic review stage (e.g., search procedure, selection, and critical appraisal); the quality of the meta-analysis methods, including the methods used for synthesizing the results, exploring heterogeneity and additional analyses, such as publication bias assessment; the significance for research use, which includes evaluating the ways authors present findings to increase accessibility and relevance for practitioners (e.g., stakeholder engagement in the process, reporting of results). We will identify strengths and weaknesses of current meta-analysis practices and compare them across time, publication outlet, and funding status.

Keywords: *meta-analysis; systematic review; methodological quality; transparency; use of research evidence*

*Correspondence: Marta Pellegrini, e-mail: marta.pellegrini@unica.it

Citation: 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. http://doi.org/10.23865/njsre.v2.6169

Highlights

- Education meta-analyses should use high-quality methods to provide reliable evidence for practice and policy decision-making.
- New approaches have been developed to improve the validity, transparency, and reproducibility of systematic review and meta-analysis methods.
- There is a need to examine the extent to which current practices in education meta-analyses align with international methodological standards.
- There is a need to examine the extent to which meta-analysis reports include information to facilitate the use of research evidence in practice.

Background

Systematic reviews with meta-analysis (hereafter meta-analyses) are transparent and comprehensive methods to synthesize quantitative research on a specific topic to estimate, for example, the effectiveness of interventions or the association between two variables. The statistical methods for meta-analysis estimate an average effect size and explore heterogeneity across studies to understand variation in studies' effects (Cooper et al., 2019).

Since the introduction of meta-analysis (Glass, 1976), interest in using this method has grown exponentially in several fields, including education. A search in ERIC conducted in September 2023 with *meta-analysis*, *meta analysis*, or *meta-analytic* as search terms yielded a total of 493 results from 1976 to 1986, 872 from 1987 to 1997, 1,815 from 1998 to 2009, and 4,822 after 2010. The reason for this interest is that high-quality meta-analyses are useful to both researchers and practitioners. For researchers, meta-analyses can synthesize existing studies and identify gaps that need further exploration. For practitioners, meta-analyses can support decision-making by providing evidence of what works under what conditions (Pigott & Polanin, 2020; Slavin, 2008). Driven by the benefit of meta-analysis, methodologists have developed systematic review and meta-analytic methods to ensure transparent, reproducible, and valid review processes and results. Tools exist to appraise the quality of the systematic review process (e.g., AMSTAR, Shea et al., 2017) as well as published methodological guidance to conduct and report both the systematic review process and meta-analysis modeling strategies (e.g. Alexander, 2020; Page et al., 2021; Pigott & Polanin, 2020). New meta-analysis approaches and associated software are also in continuous development to overcome the limits of traditional meta-analysis methods (e.g., Pustejovsky, 2020; Tipton et al., 2019b).

Despite the remarkable growth in systematic reviews, many meta-analyses published in education are often poorly conducted and their methodological quality is questionable, producing misleading results for practitioners. As an example, Tipton et al. (2022) recently compared two meta-analyses that evaluated the effectiveness of growth mindset interventions. The two reviews used different methods reaching

dissimilar conclusions about the effectiveness of the same intervention. Low quality reviews may have consequences for the use of evidence in practice, making it difficult for practitioners and policymakers to recognize what is – and what is not – strong evidence to make decisions.

Given the growth of published reviews and their impact on practice as well as the concerns about the validity of methods used in those reviews (Tipton et al., 2019a), this study aims to assess the methodological quality of meta-analyses published in education and explore how the reports addressed information to support the use of evidence in practice.

Meta-analyses and the use of evidence in practice

Considering the use of research evidence is increasingly important for educational research, and more specifically, evidence synthesis research. Meta-analyses are an important resource for promoting the use of research evidence in education, as they synthesize an entire body of evidence that may otherwise be unwieldly to practitioners and policymakers (Davies, 2000). Yet findings from prior studies have suggested practitioners and policymakers rarely use the results of meta-analyses in their work. As one approach to closing this gap, researchers might rethink what is meant by research quality in ways that also reflect the practitioners' perspectives (Mills et al., 2020; Ming & Goldenberg, 2021) and consider ways to involve stakeholders in the evidence synthesis process. In this respect, several initiatives (e.g., Rethinking Research for School or Research-Practice Partnerships) and papers (e.g., Day et al., 2023; Ming & Goldenberg, 2021; Welsh et al., 2021) have started to emphasize the role of evidence use in practice as a key point in meta-analysis and evidence synthesis processes. To examine the potential of the meta-analysis to contribute to practice, we will focus on two aspects of how the systematic review presents results. One aspect concerns the involvement of stakeholders in the conduct of the review. We will examine if the meta-analysis mentions participation of practitioners or other stakeholders in any part of the production of the systematic review. The second aspect will focus on the reporting and interpretation of the results. In order to apply meta-analysis results to practice, stakeholders need to understand how well the evidence base resembles their own context, and what the statistical results actually indicate in practical terms. We will examine how transparently the systematic review reports on the context, interventions, participants, and intervention implementation, and how the review describes the statistical results. To our knowledge, no previous meta-reviews in education evaluated how the information reported in meta-analyses may be relevant for practice.

Previous meta-reviews on the quality of meta-analyses in education

Previous studies have been conducted to evaluate the methodological and reporting quality of reviews in different disciplines. In health care and psychology this topic has been widely explored with meta-reviews focused more generally on methodological

quality (e.g., Pölkki et al., 2014), statistical techniques (e.g., Cafri et al., 2010; Linden & Hönekopp, 2021), and transparency and reproducibility (e.g., Polanin et al., 2020), showing that several areas of the review process need improvement. Within the field of education, prior research has primarily examined methodological quality by focusing on specific dimensions, such as limiting reviews to a particular type of intervention or a certain stage of the review process. We present meta-reviews conducted on methodological quality of meta-analyses in order of publication date. Table 1 provides details about each meta-review with a focus on the topic, scope, dimensions assessed, and tool used.

Valentine et al. (2009) examined 12 research syntheses on the impact of after-school programs with the aim of comparing the quality of procedures used in these reviews. More specifically, they assessed how the reviews align with best practices for research synthesis and if the inferences drawn from the included studies and results by the authors are plausible based on the evidence. The meta-review showed that the research syntheses used different conceptual and operational definitions of "after-school programs" (ASP), leading to the inclusion of different sets of studies. Criteria for inclusion also varied with some of the reviews selecting both quantitative and qualitative study designs. The authors noted that since the research question addressed by the reviews was "Do ASPs improve participants' academic achievement and/or socio-emotional well-being?" the eligibility criteria for these reviews should have focused on studies that can estimate the causal connection between ASP programs and student outcomes. However, most of the reviews narratively summarized qualitative and quantitative results. In addition to the evaluation of the quality of the review process (e.g., searching, coding), a focal point of this meta-review was to assess if the inferences drawn from the included studies and results are plausible based on the eligible studies. The authors claimed that "none of the synthesists were willing to confine their conclusions from the accumulated research to the simple statement that the evidence to date revealed nothing about the effects of ASPs, positive, negative, or null" (Valentine et al., 2009, p. 32), concluding that this may have consequences on the overgeneralization of the findings and may create confusion among policy-makers and practitioners who could make different decisions depending on the review they encounter.

Ahn et al. (2012) conducted the most comprehensive meta-review of meta-analyses with the aim to assess the methodological quality of the review process and the meta-analysis methods used. The authors searched studies mainly in journals affiliated with the American Educational Research Association (AERA) and examined 64 meta-analyses. The following stages of the review process were evaluated: problem formulation; data collection; data evaluation; data analysis; report results. Overall, the results suggested the need for a better description of the research question to include a rationale for potential moderators of the effect size. On data collection, most reviews provided sufficient information about searching, selection, and coding. Ahn et al. also noted that information about coders' agreement should be reported in future studies. On data evaluation, half of the reviews did not assess the quality of

Table 1. Characteristics of previous meta-reviews on quality of meta-analyses in education

Reference	N. reviews	Topic	Scope	Dimensions assessed	Tool used	Journal limit	Timeframe
Valentine et al. (2009)	12	Effects of after-school programs	(a) How methods used conformed to what is defined as best practice for research synthesis (b) If the inferences drawn from the included studies and results by the authors are plausible based on the data they had	(a) Conceptual and operational definitions (b) Contextualization in the syntheses (c) Sources of research literature and search terms (d) Coding reliability (e) Correspondence between the methods and implementation of individual studies and the desired inferences of syntheses (f) Summarizing and integrating the evidence from individual studies (g) Interpreting cumulative evidence (h) Consensus among the synthesis conclusions	Cooper's (2010) checklist	None	Until 2009
Ahn et al. (2012)	56	Methodological quality of the review process	(a) Examine the extent to which the current practices in meta-analyses in education are free from threats to validity (b) Guide future researchers to produce valid evidence from research syntheses	(a) Problem formulation (b) Data collection (c) Data evaluation (d) Data analysis (e) Report results	Created by the authors based on Cooper's (1982) evaluative checklist, the MUTOS framework, and MARS	Journals affiliated with AERA, Journal Educational Psychology, British Journal of Educational Psychology	2000–2010
Polanin et al. (2016)	81	Publication bias	(a) The observed difference in mean effect size between published and unpublished studies in meta-analyses	(a) Difference in effects between publication status	Comparing mean effect sizes	Review of Educational Research, Psychological Bulletin	1986–2013

Reference	N. reviews	Topic	Scope	Dimensions assessed	Tool used	Journal limit	Timeframe
Tipton et al. (2019a)	64	Meta-regression practices	(a) Current practices in light of the technical, conceptual, and practical areas of methodological development identified in Tipton et al. (2019b)	Technical characteristics: (a) methods for moderator analysis (b) test statistics (c) method to address effect size dependence Conceptual characteristics: (a) number of predictors in meta-regression (b) decision on model building (c) method to address multiplicity Practical characteristics: (a) method to address missing data (b) technique to handle missing data (c) statistical software (d) availability of data	Created by the authors based on Tipton et al. (2019b)	Previous reviews, Psychological Bulletin, Journal of Applied Psychology, Review of Educational Research, Cochrane Library	1976–2016
Hew et al. (2021)	19	Effects of flipped classroom	(a) To examine the methodological features of the meta-analyses. (b) To discuss the key methodological concerns of the meta-analyses.	(a) Sources of primary study (b) Primary study selection and data extraction (c) Primary study quality (or bias risk) (d) Issues of missing data (e) Non-independent effect sizes (f) Participant initial equivalence (g) Instructor equivalence (h) Fixed- vs. random-effects model (i) Publication bias	Created by the authors	Database search, Google Scholar	Until 2019

(*Continued*)

Table 1. (*Continued*)

Reference	N. reviews	Topic	Scope	Dimensions assessed	Tool used	Journal limit	Timeframe
Nelson et al. (2022)	22	Methodological quality of the review process	(a) Overall quality of reporting in meta-analyses (b) Frequency with which each meta-analysis QI is met across meta-analyses	Quality of reporting: (a) Research questions (b) Eligibility (c) Search procedures (d) Screening (e) Coding procedures (f) Participant information (g) Data analysis (h) Results (i) Study average	Created by the authors based on Talbott et al. (2018) quality indicators (QIs)	Database search, special education journals	2000–2020
Park et al. (2023)	29	Four best practices in meta-analytic methods	(a) Four best practices in meta-analysis	(a) Moderator analysis selection (b) Small sample corrections (c) Handling dependent effect sizes (d) Publication bias assessment	Created by the authors based on Tipton et al. (2019b)	Database search, special education journals (same set of studies as Nelson et al. (2022)	2000–2022

primary studies. Studies that included codes for study quality also used these codes to check the sensitivity of results to study quality. On data analysis, publication bias was tested only in 52% of the reviews using a variety of methods. Approximately 62% of meta-analyses mentioned the issue of effect size dependence, mostly averaging the effect sizes within studies to deal with this dependence. Of the studies reporting the choice of models, 49% used random-effects or mixed-effects models. Only five meta-analysis using mixed-effects models reported between-study variance as heterogeneity measure, while the majority reported *Q*-statistics. Information related to statistical models was often not reported in the papers, making it difficult for the authors to code some characteristics.

Polanin et al. (2016) examined publication bias, the bias that occurs when statistically significant results are more likely to be published than null results. The authors searched for studies published in *Review of Educational Research* and *Psychological Bulletin* and included 81 meta-analyses. Results showed that published studies had a mean effect size of 0.18 (SMD) higher than unpublished studies. Reviews whose only source of unpublished studies were dissertations found an insignificant difference between unpublished dissertations and published studies in their effect size.

Tipton et al. (2019a) reviewed the practices in meta-regression used in education, psychology, and medicine by focusing on how recent advanced meta-analysis methods have been applied in the literature. After searching for studies published in *Psychological Bulletin, Journal of Applied Psychology, Review of Educational Research,* and the *Cochrane Library,* 64 meta-analyses were examined. Three areas of methodological development and several related aspects were examined:

- Technical characteristics: framework for moderator analysis (subgroup, ANOVA, meta-regression); test statistics used (including new corrections for small samples); how authors addressed the issue of dependent effect sizes (ways to include one synthetic effect size or methods to model multiple effect sizes per study).
- Conceptual characteristics: for studies using multiple meta-regression, the number of predictors in the model; authors' decision on model building; how they addressed the issue of multiplicity of hypothesis tests.
- Practical characteristics: how authors addressed missing data, techniques used to handle missing covariates, statistical software used for the analysis, availability of data.

The results on technical characteristics showed that most meta-analyses relied on simple meta-regression for continuous moderators rather than multiple regression models that include both continuous and categorical moderators. In other words, many meta-analyses used models with only a single moderator at a time. Meta-analyses that did include multiple predictors failed to give a rationale for model building, lacking a clear indication of how the synthesis authors decided what moderators to include, and whether these analyses were decided *a priori* or were exploratory.

Most meta-analyses used default test statistics that assume a normal sampling distribution, without adjustments for small samples. Although the included reviews had an average of 4.5 effect sizes per study, few reviews used techniques to account for the dependence among effect sizes within studies. Finally, corrections for multiple comparisons were rarely used. The results on practical characteristics showed that the issue of missing data was often mentioned in syntheses and usually addressed with ad hoc approaches, such as list-wise deletion or separate analyses for each moderator. On software, only 11% of the syntheses used R packages, while most used "point-and-click" software that does not allow for more complex analyses. Many reviews shared data in accessible formats, such as tables in appendices or in online repository, but often in a form of partial dataset including selected information.

Hew et al. (2021) examined methodological features of 19 meta-analyses conducted on the impact of flipped classrooms on student outcomes, and discussed key methods concerns to inform practitioners about the trustworthiness of the results. The findings of the analysis showed that most of the meta-analyses used high-quality procedures for searching, selecting, and coding studies, although strategies for gray literature search could be improved. The following features resulted as the main critical issues to address in future reviews on flipped classrooms: assessing risk of bias of primary studies; establishing initial equivalence between groups on key characteristics; addressing effect sizes dependence within studies; assessing publication bias with methods other than funnel plot.

Nelson et al. (2022) evaluated the reporting quality of 22 meta-analyses on mathematics interventions for students with or at risk of disabilities, focusing on the quality of review procedures and data analysis, using a tool created by Talbott et al. (2018). The average quality score was 61%, with a range of 47% to 80% across meta-analyses. The categories with the highest mean scores were research questions (95%) and data analysis (77%), while participant demographics (39%) and screening (48%) had the lowest mean scores. Based on their quality ratings, Nelson et al. suggest that inclusion criteria should specify all characteristics relevant for study inclusion or exclusion (e.g., PICOS framework); the credentials or expertise of those conducting the literature search should be reported; and details about screening reliability and training of reviewers to conduct screening should be included. The authors suggested using the PRISMA guidelines for reporting the screening process. On coding procedures, the authors noted the need for making codebooks accessible, as well as the need for more information on participant demographic characteristics. The lack of details about participants' characteristics in eligible studies may impact the ability to correctly generalize findings. Reporting on data analysis and methods used could also be improved.

Park et al. (2023) further explored a similar set of studies (29 studies) by Nelson et al. (2022) with the aim of evaluating four practices on which the literature on meta-analysis appears to have a robust methodological consensus: moderator analysis selection (subgroup, ANOVA, meta-regression); small-sample corrections (statistical test used); handling dependent effect sizes (ways to report a single effect per study,

multilevel/multivariate meta-analysis models, use of Robust Variance Estimation – RVE); and methods for publication bias assessment. On moderator analysis, most meta-analyses used subgroup analysis or simple meta-regression, with more recent studies (2011–2022) tending to use meta-regression with multiple moderators. Most studies (65%) neither mentioned nor used a small-sample correction. Most of the meta-analyses failed to include multiple effect sizes per study, opting for choosing a single effect size or averaging effect sizes within studies. More recent meta-analyses tended to use RVE. There is an increasing trend in assessing publication bias, although 48% of the studies used funnel plots.

Wang et al. (2023) presented an ongoing meta-review on the use of risk of bias assessments in educational meta-analyses at the annual meeting of the Society for Research on Educational Effectiveness. The preliminary results of 87 included studies showed that many reviews did not perform a critical appraisal of primary studies, and when conducted, many tools were created by the authors.

Our meta-review builds on the findings of the studies discussed above to provide a more comprehensive and up-to-date meta-review on education meta-analyses similar to Ahn et al. (2012). The proposed meta-review also considers a broader set of meta-analyses than previous meta-reviews, focusing on the impact of educational interventions in K–12 on student academic achievement, with no restriction regarding the publication outlet. Prior meta-reviews restricted their searches to journals with high-impact factors (such as those published by the American Educational Research Association). This meta-review also assesses both the quality of the systematic review process (e.g., problem formulation, searching, reproducibility) and the meta-analysis methods (e.g., synthesis methods, heterogeneity). Finally, the proposed meta-review explores a novel area: how meta-analysts both directly (via stakeholder engagement) and indirectly (via data visualization and presentation of findings) incorporate strategies for making their meta-analyses useful for non-research audiences.

Aim and research questions

The main purpose of this meta-review is to evaluate the methodological quality of systematic reviews with meta-analysis published in the field of education and review the reporting of results to facilitate the use of research evidence in practice. Three major dimensions of meta-analyses will be assessed considering both the degree to which the researchers reported these characteristics and if they used current best-practice meta-analysis methods (e.g., Page et al., 2021; Pigott & Polanin, 2020; Tipton et al., 2019a): (i) the quality of the *review process*, related to the procedures and standards used to conduct the systematic review (e.g., search procedures) as well as the reproducibility of the process (e.g., protocol preregistration, the availability of data); (ii) the quality of the *meta-analysis methods*, 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); (iii) the

significance for research use, related to the inclusion of analyses, results, and interpretations that support the use of the evidence in practice (e.g., under which conditions the intervention works, section on the implications for practice).

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

- To what extent are best-practice systematic review process methods used and reported in meta-analyses in education?
- To what extent are best-practice meta-analysis methods used and reported in meta-analyses in education?
- How do the systematic review and meta-analysis methods used in education meta-analyses differ across publication outlet, funding status and across time?
- To what extent and in which ways have researchers used strategies that support the use of research evidence in practice in education meta-analyses?

Methods

Inclusion criteria

Meta-analyses will be included if they meet the following criteria:

- The meta-analysis includes general populations of students in K–12. We include pre-K or post-secondary only when other K–12 grades are included. We exclude meta-analyses solely focused on special education populations (including learning disabilities).
- The meta-analysis focuses solely on school-based academic interventions. We exclude 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. We include motivation interventions if they are focused on academic achievement.
- The meta-analysis reports a summary effect size for student academic achievement. We exclude other education-related outcomes, such as socio-emotional skills, attendance, dropout rates, computational thinking, and teacher outcomes.
- The meta-analysis includes studies using group designs (i.e., randomized controlled trials, quasi-experimental designs). We exclude 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. We exclude these other design types as we are most interested in meta-analyses that include study designs that can best support a causal inference about the effectiveness of the academic intervention.
- The meta-analysis is published in English between January 2011 and September 2023. The latest comprehensive review on the quality of systematic reviews included studies published until 2010 (Ahn et al., 2012). Furthermore, Tipton et al. (2019a) indicated 2010 as the beginning of a new phase of methodological growth in the field of meta-analysis.

- The meta-analysis is published in a peer-reviewed journal. We restrict our search to peer-reviewed papers excluding gray literature for two reasons. Peer-reviewed studies have already passed an expert evaluation for their quality. Some unpublished studies (e.g., conference papers) may not report the full method sections. For a similar equitable reason, we exclude dissertations and reports due to their typically longer page count compared to articles.

Search strategy and sources

We will identify studies for the current review using three search strategies. First, we will search databases including Academic Search Ultimate, APA PsycINFO, ERIC, Teacher Reference Center via EBSCO and Social Sciences Citation Index, Science Direct. Categories of keywords related to meta-analysis, intervention study, and participant will be used to search for studies and will be adapted as needed depending on the database. Draft search strings are reported in Table 2. Second, we will hand search the tables of contents of journals that published mainly research syntheses using *Paperfetcher* (Pallath & Zhang, 2023): *Review of Educational Research, Educational Research Review, Review of Research in Education, Journal of Research on Educational Effectiveness, Campbell Systematic Reviews.* Finally, we will check the reference list of all meta-analyses included in previous meta-reviews (see Supplementary materials) with the aim of analyzing the quality of methodological features.

Table 2. Draft 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) ANDALL (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

Selection procedure

Following the identification of potential studies, a two-stage process will be used for selection. First, the title and abstract of the located studies will be single screened and only records that are meta-analyses in K-12 education will be retained for full-text screening. Next, the full text of the retained studies will be reviewed based on the inclusion criteria. Studies will be reviewed in double screening by two independent reviewers to avoid the exclusion of potential eligible studies. The reviewers will meet regularly to resolve conflicts and improve the level of agreement. The study selection process will be carried out using *Covidence* (https://www.covidence.org/), an online platform for reviews that supports a more systematic and transparent process.

Codebook and coding procedure

The codebook will consist of the three main dimensions under evaluation: the review process, the meta-analysis methods, and the significance for research use. We developed a draft of the first two dimensions based on existing tools for review quality assessment (e.g., AMSTAR, MARS), guidelines (Page et al., 2021; Pigott & Polanin, 2020), and codebooks from previous educational meta-reviews (see Supplementary materials). We started by listing all items from existing tools and codebooks from previous meta-reviews, as well as the key points outlined in Pigott and Polanin's (2020) guidelines. We grouped them based on the aspects assessed (e.g., information sources) and the review stage (e.g., searching). We then combined items from different tools that addressed the same aspect, ensuring to retain all relevant details assessed by each tool. After creating all items of our codebook based on this procedure, we organized them in three sections according to the characteristics assessed: (i) *Background information* of the meta-analysis (e.g., number of studies, journal; 11 items); (ii) *Review process* with the following dimensions – problem formulation (2 items), inclusion criteria (7 items), searching (3 items), selection (6 items), coding (4 items), critical appraisal (4 items), reproducibility (4 items); (iii) *Meta-analysis methods* with the following dimensions – synthesis (16 items), heterogeneity (5 items), additional analysis (8 items).

On the *Significance for research use*, all meta-analyses included in our meta-review from 2021 will be coded to assess the extent to which authors (i) explicitly included stakeholders in the research process or discussed how findings can be used by stakeholders, (ii) reported findings that support the application of the interventions in similar contexts, and (iii) made accessible the findings to stakeholders. A growing body of literature highlights the importance of involving stakeholders in the systematic review process (e.g., Harris et al., 2016). Incorporating stakeholder voices not only increases the relevance of findings to meet stakeholder needs but may also be the "key to ethical decision making, which is the only sustainable solution to inequities" (Cellier, 2021). Even when researchers are not able to directly involve stakeholders in the research process, there are ways to make findings more accessible and relevant for stakeholders. These approaches include reporting student, school, and intervention

characteristics so that stakeholders can assess how well the intervention might fit with their local context (Farley-Ripple et al., 2018). In drafting our codebook, we used the PICOTS framework to code which characteristics of participants, interventions, comparisons, outcomes, times, and settings were considered by the authors and were tested as potential moderators of the effect. We will also code if the authors provided data visualizations or effect size transformations to support stakeholders' understanding of findings (e.g., Fitzgerald & Tipton, 2023). The codebook will include the following groups of items: stakeholder engagement (2 items); reporting of results (14 items); interpretation of results for broader audience (6 items).

The draft codebook (see Supplementary materials) will be piloted by all coders on several studies to be revised as necessary. A set of ten studies will be used to train coders and to reach a high level of agreement. For the *Review process*, MP will code all studies and NP and HS will double-code a 25% of the studies. For the *Meta-analysis methods*, MP and TP will single-code all studies and CC will double-code 25% of the studies. For the *Significance for research use* coding form, ED and HS will code all of the studies and MP will double-code 25% of the studies. Regular meetings will be conducted to check on coding. Data extraction will be conducted through *MetaReviewer* (https://www.metareviewer.org) that allows the use of a systematic and transparent process.

Data analysis

We will analyze studies descriptively by providing descriptive statistics for the characteristics coded and an average quality score for the three dimensions assessed. We will calculate the frequency and percentage across the options for categorical characteristics. We will examine the mean, standard deviation, minimum, and maximum for continuous characteristics. If a study reports more than one meta-analysis (e.g., different outcomes), we will consider all of them in the background information section, but we will code the methods characteristics used in the first reported meta-analysis.

For the *Review process* and *Meta-analysis methods* sections, we will compare the differences in methodological aspects across time, journal, and funding agency. We hypothesize that meta-analyses recently published would report a larger number of methodological characteristics as well as use modern meta-analytical methods. We expect that meta-analyses published after 2020 would incorporate a larger number of best practices that have reached methodological consensus, such as methods to deal with dependent effect sizes, and publication bias assessment. Based on previous studies (e.g., Tipton et al., 2019a) we expect a few studies to consider more sophisticated methods which, despite gaining consensus among statisticians, have not yet been integrated into current practices. Among them, we expect a few studies to use small-sample corrections, techniques to address multiplicity, and principled methods to handle missing data. We hypothesize that meta-analyses published in journals with a higher impact factor and devoted to research syntheses (e.g., *Review of Educational*

Research, Educational Research Review) would have a higher quality score in characteristics related to the review process and the meta-analytical methods. This is because we expect that reviewers for these journals have a stronger methodological expertise compared to those reviewing for journals primarily focused on educational content. We hypothesize that funded meta-analyses (especially from organizations such as the Institute of Education Sciences and the Education Endowment Foundation) would have a higher quality than unfunded ones, given the intense competition to secure funding. All analyses, figures, and tables will be conducted using the R software.

Author biographies

Marta Pellegrini is assistant professor in research methods in education at the University of Cagliari. Her research interests focus on evidence-based education, including systematic reviews with meta-analysis, and impact evaluations on educational interventions.

Terri Pigott is a Professor at the College of Education and Human Development, Georgia State University. Dr. Pigott's 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 doctoral student in research, measurement, and statistics at the College of Education and Human Development, Georgia State University. Her research focuses on methodological advances, best practices, and the policy implications of meta-analysis.

Elizabeth Day is a research assistant professor at the HEDCO Institute for Evidence-Based Educational Practice, University of Oregon. Her research focuses on understanding best practices for connecting research, practice, and policy. Her particular area of focus is on the use of evidence syntheses in social policy and educational practice. Day is also visiting lecturer for Cornell University's School of Public Policy where she teaches a course on translational research.

Natalie Pruitt is completing her master's degree in educational research, measurement, and statistics at Georgia State University. Her research interests include education, families, assessments, and community-based programs.

Hannah F. Scarbrough is a PhD student in education policy studies and a graduate research assistant at Georgia State University. Her research interests are in systematic review and meta-analysis research methods, research use and dissemination for policymaking, and wrap-around school and support services.

References

Alexander, P. A. (2020). Methodological guidance paper: The art and science of quality systematic reviews. *Review of Educational Research*, *90*(1), 6–23. https://doi.org/10.3102/0034654319854352

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

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research*, *45*(2), 239–270. https://doi.org/10.1080/00273171003680187

Cellier, M. S. (2021). The place and importance of patients in deliberative processes. *Frontiers in Medical Technology, 3*. https://doi.org/10.3389/fmedt.2021.794695

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

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52*, 291–302. https://doi.org/10.3102/0034654305200229

Cooper, H. (2010). *Research synthesis and meta-analysis*: *A step-by-step approach*. Sage Publications.

Davies, P. (2000). The relevance of systematic reviews to educational policy and practice. *Oxford Review of Education, 26*(3–4), 365–378. https://doi.org/10.1080/713688543

Day, E., Shimmel, L., Steynka-Fry, K., Tanner-Smith, E., & Grant, S. (2023). *Using systematic reviews to inform practice: Engaging with educators* [Paper presentation]. Annual Meeting of the Society for Research and Reform in Education 2023, Arlington, VA.

Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher, 47*(4), 235–245. https://doi.org/10.3102/0013189X18761042

Fitzgerald, K. G., & Tipton, E. (2023). A knowledge mobilization framework: Toward Evidence-based statistical communication practices in education research. *Journal of Research on Educational Effectiveness*, 1–21. https://doi.org/10.1080/19345747.2023.2209082

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

Hallinger, P. (2022). Mapping the contributions of the Review of Educational Research to education, 1931–2020. *Review of Educational Research, 93(*6), 791–824. https://doi.org/10.3102/00346543221132771

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

Harris, J., Croot, L., Thompson, J., & Springett, J. (2016). How stakeholder participation can contribute to systematic reviews of complex interventions. *Journal of Epidemiology and Community Health, 70*(2), 207–214. https://doi.org/10.1136/jech-2015-205701

Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, *16*(2), 358–376. https://doi.org/10.1177/1745691620964193

Mills, K. J., Lawlor, J. A., Watling Neal, J., Neal, Z. P., & McAlindon, K. (2020). What is research? Educators' conceptions and alignment with United States federal policies. *Evidence & Policy*, *16*(3), 337–358.

Ming, N. C., & Goldenberg, L. B. (2021). Research worth using: (Re) framing research evidence quality for educational policymaking and practice. *Review of Research in Education*, *45*(1), 129–169. https://doi.org/10.3102/0091732X21990620

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

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J., Akl, E., Brennan, S., Chou, E., Glanville, J., Grimshaw, J., Hrobjartsson, A., Lalu, M., Li, T., Loder, E., 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

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

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., 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

Pölkki, T., Kanste, O., Kääriäinen, M., Elo, S., & Kyngäs, H. (2014). The methodological quality of systematic reviews published in high-impact nursing journals: A review of the literature. *Journal of Clinical Nursing*, *23*(3–4), 315–332. https://doi.org/10.1111/jocn.12132

Pustejovsky, J. (2020). clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections (Version R package version 0.4.1) [Computer software]. https://CRAN.R-project.org/package=clubSandwich

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*. https://doi.org/10.1136/bmj.j4008

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5–14. https://doi.org/10.3102/0013189X08314117

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*, 161–179. https://doi.org/10.1002/jrsm.1338

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

Talbott, E., Maggin, D., Van Acker, E. Y., & Kumm, S. (2018). Quality indicators for reviews of research in special education. *Exceptionality*, *26*, 245–265. https://doi.org/10.1080/09362835.2017.1283625

Valentine, J. C., Cooper, H., Patall, E. A., Tyson, D., & Robinson, J. C. (2010). A method for evaluating research syntheses: The quality, conclusions, and consensus of 12 syntheses of the effects of after-school programs. *Research Synthesis Methods*, *1*(1), 20–38. https://doi.org/10.1002/jrsm.3

Wang, J., Wang, X., Zhang, Q., Bao, H., Zhao, Z., & Neitzel, A. (2023). Risk of bias tools in systematic reviews of interventions PK-12 education: A critical analysis of studies in Review of Educational Research [Paper presentation]. Annual Meeting of the Society for Research on Educational Effectiveness 2023, Arlington, VA. https://sree.confex.com/sree/2023/meetingapp.cgi/Paper/4792

Weise, A., Büchter, R., Pieper, D., & Mathes, T. (2020). Assessing context suitability (generalizability, external validity, applicability or transferability) of findings in evidence syntheses in healthcare—An integrative review of methodological guidance. *Research Synthesis Methods, 11*(6), 760–779. https://doi.org/10.1002/jrsm.1453

Welsh, R. O. (2021). Assessing the quality of education research through its relevance to practice: An integrative review of research-practice partnerships. *Review of Research in Education, 45*(1), 170–194. https://doi.org/10.3102/0091732X20985082

Supplementary materials

List of previous meta-reviews

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.

Hallinger, P. (2022). Mapping the Contributions of the Review of Educational Research to Education, 1931–2020. *Review of Educational Research*, 00346543221132771.

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.

Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, *16*(2), 358–376.

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.

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*, 07319487231185133.

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.

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.

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.

Wang, J., Wang, X., Zhang, Q., Bao, H., Zhao, Z., Neitzel, A. (2023). Risk of Bias Tools in Systematic Reviews of Interventions PK-12 Education: A Critical Analysis of Studies in Review of Educational Research. Paper presented at the Annual Meeting of the Society for Research on Educational Effectiveness 2023, Arlington, VA. Retrieved from https://sree.confex.com/sree/2023/meetingapp.cgi/Paper/4792

List of existing guidelines and tools to assess the quality of meta-analysis

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.

American Psychological Association Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: why do we need them? What might they be? *Am. Psychol.* 63, 839–851. https://doi.org/10.1037/0003-066X.63.9.839

Aromataris, E., Fernandez, R., Godfrey, C., Holly, C., Kahlil, H., Tungpunkom, P. (2015). Summarizing systematic reviews: methodological development, conduct and reporting of an Umbrella review approach. *Int J Evid Based Healthc, 13*(3), 132–140.

Critical Appraisal Skills Programme UK. (2013). *CASP checklists*. Retrieved from http://www.casp-uk.net/#!casp-tools-checklists/ c18f8

Higgins, J. P., Lane, P. W., Anagnostelis, B., Anzures-Cabrera, J., Baker, N. F., Cappelleri, J. C., … & Whitehead, A. (2013). A tool to assess the quality of a meta-analysis. *Research Synthesis Methods*, *4*(4), 351–366.

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.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic reviews*, *10*(1), 1–11.

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*, 07319487231185133.

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.

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.

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.

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.

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez-Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., … & López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-analyses: The REGEMA checklist. *Research Synthesis Methods*, *12*(4), 516–536.

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., … & 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*.

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.

Draft codebook

Category	Item type	Item	Options
Background information			
Study information	Free text	Authors	
	Free text	Year of publication	
	Free text	Journal of publication	
	Dropdown	Was there funding received to support this study?	Yes, No, Not mentioned
	Free text	If yes, report the name of the organization that funded the study	
	Dropdown	Is there a conflict of interest with this report?	Yes, No, Not mentioned
	Free text	If yes, what is the nature of the COI?	
	Free text	The type of interventions studies	
	Dropdown	The outcomes evaluated	Language art, Reading, Mathematics, Science, Social studies, Academic achievement in general, Other (describe)
	Free text	Number of studies included in the review	
	Free text	Number of effect sizes included in the review	
Review process			
Problem formulation	Dropdown	Did the authors provide an explicit research question/objective on estimating the average treatment effect?	Yes, No
	Dropdown	Did the authors provide an explicit research question/objective on exploring the variation of the average effect across studies?	Yes, No
Inclusion criteria	Dropdown	Did the authors list inclusion criteria?	Yes, No
	Dropdown	Did the authors list inclusion criteria for the population of interest?	Yes, No
	Dropdown	Did the authors list inclusion criteria for eligible interventions?	Yes, No
	Dropdown	Did the authors list inclusion criteria for the comparison or control condition?	Yes, No
	Dropdown	Did the authors list inclusion criteria for eligible outcomes?	Yes, No
	Dropdown	Did the authors list inclusion criteria for eligible study designs?	Yes, No
	Checkbox	Did the authors list other inclusion criteria for eligibility of studies or reports?	Publication type, Language, Timeframe, No, Other (describe)

(*Continued*)

Category	Item type	Item	Options
Searching	Checkbox	Which types of search strategies did the authors use?	Database search, Hand search journals, Retrospective reference harvesting, Prospective forward citation searching, Search engine, Contacting authors, Other (describe)
	Checkbox	Did the authors specifically search for gray literature?	Dissertations and/or theses, Contacting researchers, Searches of websites of independent research firms (i.e., AIR, Rand, Mathematica), Searches of websites of related associations (i.e., Autism advocacy groups), No, Other (describe)
	Dropdown	Are search terms given for all database searches?	Complete strings for each database searched, Complete strings for some databases searched, Example/ list of terms, No
Selection	Dropdown	Did the authors use a two-stage screening process?	Yes, No, Not mentioned
	Dropdown	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
	Dropdown	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
	Checkbox	Di the authors report other procedures to check the accuracy of the selection stage?	IRR, Describe a consensus process, State that conflicts were resolved without describing the consensus, Not mentioned, Not applicable
	Dropdown	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)
	Dropdown	Did the authors report a PRISMA flowchart?	Yes, Incompletely reported, Not reported

Category	Item type	Item	Options
Coding	Checkbox	How did the authors report information about the characteristics coded?	Codebook, Narrative description of the characteristics, Not reported
	Dropdown	Which approach did the authors use for coding?	Single-screening, Single-screening with validation from a second screener, Partial double-screening, Independent double-screening, Not mentioned
	Checkbox	Di the authors report other procedures to check the accuracy of coding?	IRR, Describe a consensus process, State that conflicts were resolved without describing the consensus, Not mentioned, Not applicable
	Dropdown	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	Dropdown	Which approach did the authors use for critical appraisal?	Front-end approach, Back-end approach, No critical appraisal
	Checkbox	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 (2015) WI for group design, Effective Public Health Practice Project, JBI institute, Newcastle-Ottawa Scale, ROBINS-I, WWC standards, Selected items by authors, Other (describe)
	Checkbox	If "selected items by authors", which characteristics did they code?	Study design, Control group type, Implementation quality, Measurement, Attrition, Baseline equivalence, Publication status, Assignment level, Other (describe)
	Dropdown	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

(*Continued*)

Category	Item type	Item	Options
Reproducibility	Dropdown	Did the authors mention pre-registration of the protocol?	Preregistered protocol in a peer-reviewed repository, Preregistered protocol in a non peer-reviewed repository, No
	Dropdown	If the authors pre-registered a protocol, did they mention any changes?	Yes, No
	Dropdown	Did the authors make the dataset available?	Yes in supplemental materials, Yes in an online repository, No, Other (describe)
	Dropdown	Did the authors make the statistical code available?	Yes in supplemental materials, Yes in an online repository, No, Other (describe)
Meta-analysis methods			
Synthesis methods	Free text	Average effect size of the main meta-analysis	
	Checkbox	Which measure of uncertainty did the authors provide for the average effect size?	Standard error, Confidence interval, Not mentioned
	Free text	Standard error of average effect size of the main meta-analysis (if provided)	
	Free text	Confidence interval of average effect size of the main meta-analysis (if provided)	
	Checkbox	How did the authors calculate study effect sizes?	Unadjusted effect size, adjusted effect size, Not mentioned
	Dropdown	Did the authors mention they included studies with clusters as the unit of assignment?	Yes, No
	Dropdown	If yes, did the authors adjust effect sizes and variances for clustering?	Yes mentioned Cochrane Handbook, Yes mentioned Hedges et al. (2007), No
	Dropdown	Did the authors say they used fixed-effects model or random-effects model?	FEM, REM, No (describe)
	Dropdown	Did the authors mention there were dependent effect sizes within studies?	Yes, No
	Dropdown	Did the authors mention there were dependent effect sizes within studies?	Yes, No
	Dropdown	How did the authors handle dependent effect sizes?	Averaged/composite, Selected one, Subgroup approaches/shifting unit-of-analysis approach, Model-based methods, Not mentioned, Other (describe)

Category	Item type	Item	Options
	Dropdown	If using a model-based method for dependent effect sizes, which one did the authors use?	Multivariate meta-analysis, Multilevel meta-analysis, Multivariate meta-analysis/ Correlated Effects model with RVE, Multilevel meta-analysis/ Hierarchical Effects model with RVE, Correlated and Hierarchical Effects (CHE) model, CHE with RVE
	Dropdown	Which statistical test did the authors use for the mean effect size?	t-test, Permutation test, Default/not mentioned, Other (describe)
	Dropdown	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)
	Dropdown	Did the authors use a method for multiple comparisons correction?	Ad hoc methods, Benjamini-Hochberg, Bonferroni, Permutation test, No, Other (describe)
	Checkbox	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	Dropdown	Did the authors test for heterogeneity of the mean effect size?	Yes, No
	Checkbox	If yes, which measure of heterogeneity did the authors provide?	Q, I-squared, tau-squared, 95% Prediction Interval, Not reported, Other (describe)
	Checkbox	If there is heterogeneity, how did the authors model it?	Subgroup oneway, Subgroup multiway, ANOVA oneway, ANOVA multiway, Meta-regression simple, Meta-regression multiway, Not modeled, Not reported, Other (describe)
	Free text	If the authors used multiple meta-regression, how did they describe the process of model building?	
	Dropdown	Did the authors distinguish between confirmatory and exploratory analysis?	Yes, No

(*Continued*)

Category	Item type	Item	Options
Additional analysis	Dropdown	Did the authors mention publication bias as a potential issue?	Yes, No
	Checkbox	If the authors assessed publication bias, which methods did they use?	Any fail-safe N, Funnel plot, Egger's regression, Egger sandwich, Selection modeling, Unpublished vs. published as a moderator, Trim and fill method, Not mentioned, Other (describe)
	Dropdown	Did the authors mention missing data as a potential issue?	Yes, No
	Checkbox	If the authors handled missing data, which procedures did they use?	Contact the authors, Check other published syntheses/reports, Ad hoc methods (i.e., list-wise deletion), Multiple imputation, Other principled methods (FIML), After compare results fo different models, Other (describe)
	Dropdown	Did the authors mention outliers as a potential issue?	Yes, No
	Checkbox	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, Other (describe)
	Checkbox	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, 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
	Free text	Did the authors perform any other additional analysis? Describe	
Significance for research use			
Stakeholder engagement	Dropdown	Did the authors include stakeholders in the research process?	Yes, No
	Checkbox	If yes, at which stages?	Research design, Data collection, Data analysis, Interpretation of findings, Other

Category	Item type	Item	Options
Reporting of results	Dropdown	Did the authors report the sampling methods used in the included studies?	Yes, No
	Dropdown	Did the authors report a table with the characteristics of the included studies?	Yes, No
	Checkbox	Which participant characteristics did the authors report?	Grade level, Race/ethnicity, Socio-economic status, Special needs, Other
	Checkbox	Which intervention characteristics did the authors report?	Implementation fidelity, Interventionist, Instructional setting, Cost for materials, Cost for teacher training, Other
	Checkbox	Which comparison characteristics did the authors report?	Nature of the intervention, Other
	Checkbox	Which outcome characteristics did the authors report?	Measure type, Other
	Checkbox	Which time characteristics did the authors report?	Intervention duration, Intervention intensity, Outcome timing, Teacher training intensity, Other
	Checkbox	Which setting characteristics did the authors report?	Country, Urbanicity, Private vs. public school, Other
	Checkbox	Which participant characteristics did the authors test as potential moderators?	Grade level, Race/ethnicity, Socio-economic status, Special needs, Other
	Checkbox	Which intervention characteristics did the authors test as potential moderators?	Implementation fidelity, Interventionist, Instructional setting, Cost for materials, Cost for teacher training, Other
	Checkbox	Which comparison characteristics did the authors test as potential moderators?	Nature of the intervention, Other
	Checkbox	Which outcome characteristics did the authors test as potential moderators?	Measure type, Other
	Checkbox	Which time characteristics did the authors test as potential moderators?	Intervention duration, Intervention intensity, Outcome timing, Teacher training intensity, Other
	Checkbox	Which setting characteristics did the authors test as potential moderators?	Country, Urbanicity, Private vs. public school, Other

(*Continued*)

Category	Item type	Item	Options
Interpretation of results for broader audience	Dropdown	Did the authors transform effect sizes into different metrics?	Yes, No
	Checkbox	If yes, which metrics did they use?	Cohen's U3, CLE, Intent to treat, Years of Learning, Benchmarking, Percentiles, Thresholds, Other
	Dropdown	Did the authors present data visualization of findings?	Yes, No
	Checkbox	If yes, which visualizations did they use?	Forest plot, Bar Plots, Rainforest Plots, Meta-Analytic Rain Cloud Plot, other
	Dropdown	Did the authors explicitly mention the use of findings by practitioners or policymakers?	Yes, No
	Free text	Copy/paste relevant text on the use of findings by practitioners or policymakers	