

The Montessori Method Under Scrutiny: A Meta-Analytic and Metascientific Review for Greater Transparency in Education Research

Samuel P. León¹ , Anastasiya A. Lipnevich² ,
and María García-Garrido¹ 

Abstract

In this preregistered systematic review, we identified 198 studies, of which 114 provided sufficient quantitative data for meta-analysis. Across these studies, the Montessori method was associated with nominally positive and conventionally large, pooled effects on learning outcomes in pretest–posttest and posttest-only designs. However, these estimates were characterized by extreme between-study heterogeneity, substantial small-study effects, and clear indications of publication bias, which substantially limited their robustness and interpretability. Methodological limitations, including inconsistent reporting of implementation fidelity, low transparency, and limited experimental control, further constrained the generalizability of the findings. By combining design-sensitive modeling, multimethod bias diagnostics, and an operational framework for Montessori fidelity, this review advances a metascientific perspective on the credibility of evidence in education research. Overall, the results indicate that the large average effects reported in the literature are largely attributable to structural weaknesses of the evidence base rather than to robust educational impacts, underscoring the need for greater methodological rigor, preregistration, and open data to strengthen cumulative knowledge about Montessori education.

¹Departamento de Pedagogía, Universidad de Jaén, Jaén Spain

²Queens College and the Graduate Center, City University of New York, New York, NY, USA

Corresponding Author:

Samuel P. León, Departamento de Pedagogía, Universidad de Jaén, Campus de las Lagunillas, SN, Jaén, CP 23071, Spain.

Email: sparra@ujaen.es

Keywords

Montessori, interventions, learning, meta-analysis

Theoretical Introduction*Background and Rationale*

For over a century, the Montessori approach to education has captured the attention of parents, educators, and researchers alike. Developed by the trailblazing Italian physician and educator Maria Montessori in the early 1900s, this child-centered philosophy has gained widespread popularity and garnered both ardent praise and fierce criticism in nearly equal measure. Yet despite its historical significance and global spread, the evidence regarding its actual educational impact remains contested. The present review adopts a metascientific stance: Rather than advocating for or against the method, we ask what the totality of research reveals when examined transparently and systematically. Beyond the longstanding question of whether Montessori education is effective, an equally important issue concerns the credibility and robustness of the evidence on which such claims are based.

From its origins, the Montessori movement developed in dialogue, and at times in tension, with other major figures of progressive education such as John Dewey, William Kilpatrick, Ovide Decroly, Rudolf Steiner, and Célestin Freinet, who each advanced alternative conceptions of child-centered and experiential learning. Their differing perspectives on autonomy, structure, and the social function of schooling helped to shape both the diffusion and the critique of Montessori pedagogy. More recent analyses, such as L'Ecuyer (2023), have revisited these debates, emphasizing how philosophical diversity and historical controversy continue to inform the modern interpretation and evaluation of Montessori education.

Over the decades, Montessori education has expanded across more than 150 countries, spanning private and public settings (Debs et al., 2022; A. Murray et al., 2023). National reports illustrate this expansion: in Italy, the number of accredited Montessori classrooms more than doubled between 2013 and 2021 (Scippo, 2022); similar growth has occurred in specific regional contexts, such as Catalonia (Spain) (Scippo & Cañigueral-Viñals, 2023), and adaptations within the Chinese educational context illustrate increasing processes of cultural reinterpretation (Chen, 2021). Despite this broad diffusion, Montessori research has historically been unevenly distributed, dominated by studies from Western, middle-class contexts and only recently engaging with culturally diverse and under-represented communities (Canzoneri-Golden & King, 2023; Moquino et al., 2023).

When it comes to empirical findings, they remain mixed. Advocates emphasize benefits for academic performance, self-regulation, and socio-emotional development (A. Lillard & Else-Quest, 2006; A. S. Lillard et al., 2023), whereas critics highlight methodological inconsistencies and context-specific effects (Lopata et al., 2005). Such tensions underscore the need for a comprehensive, methodologically

rigorous synthesis that distinguishes genuine effects from artifacts of design or reporting. To evaluate Montessori research, it is first necessary to understand what constitutes Montessori education in pedagogical and operational terms.

The Montessori Educational Framework

Maria Montessori's educational philosophy is grounded in a set of fundamental ideas and pedagogical principles articulated throughout her writings. For the purposes of the present review, and drawing on a close reading of Montessori (2004), we analytically organized these ideas into nine core pedagogical principles that capture central dimensions of her educational approach and that could, in principle, be operationalized for the empirical examination of implementation fidelity. Additionally, Montessori described five broad curricular areas through which her educational approach is typically enacted. It is beyond the scope of the current paper to provide a detailed description of these principles and curricular areas, but we briefly outline them and refer the reader to the extensive Montessori literature for a more comprehensive exploration (Montessori, 2004).

Importantly, we do not treat this nine-principle structure as a prescriptive or canonical standard of authenticity. Rather, it represents an analytic synthesis developed for coding purposes. Montessori herself emphasized contextual adaptation to children's cultural and developmental needs. Our framework, therefore, translates these theoretical constructs into observable and codable indicators (see Appendix A), allowing fidelity to be examined empirically while acknowledging diversity within the Montessori community (Ferrero et al., 2021a; A. K. Murray & Daoust, 2023).

Of the nine analytically identified principles, eight could be operationalized at the level of study reports. One principle (absorbent mind), was deemed too broad and abstract to be reliably coded based on published descriptions and was therefore excluded from the empirical coding and subsequent analyses (see Appendix A).

The following nine pedagogical principles were analytically identified for the purposes of the present review:

1. **Freedom within limits:** Montessori emphasized that children should have the freedom to choose their activities and work at their own pace. However, this freedom exists within a structured environment, which provides clear boundaries to ensure learning.
2. **Prepared environment:** The learning environment is designed to facilitate self-directed learning. Materials are carefully selected and arranged to support independence, and the environment is adapted to the child's needs and development stage.
3. **Autoeducation (self-education):** Montessori believed that children have an innate ability to teach themselves through interaction with their environment. The role of the teacher is to guide rather than instruct, allowing children to discover knowledge through exploration.
4. **Respect for the child:** Central to Montessori philosophy is respect for each child's individuality and developmental needs. Educators provide guidance

based on observing children's interests and progress, fostering a deep sense of respect and trust in their abilities.

5. Sensitive periods: Montessori identified specific windows of opportunity during which children are particularly receptive to learning certain skills. During these sensitive periods, children can absorb new knowledge effortlessly, making it crucial for educators to recognize and support these phases.
6. Absorbent mind: In early childhood, Montessori emphasized the child's ability to effortlessly absorb information from the environment. This principle underscores the importance of a well-prepared and stimulating environment.
7. Intrinsic motivation: Montessori's approach encourages learning driven by intrinsic motivation rather than external rewards or punishments. Children are motivated by their own curiosity and desire to master new challenges.
8. Individualized learning: Each child learns at their own pace, and the Montessori method accommodates this by allowing individualized progression through the curriculum. There is no uniform standard or schedule but rather a focus on each child's unique learning trajectory.
9. Holistic development: The Montessori method aims for the holistic development of the child, fostering growth in physical, social, emotional, and cognitive domains. This holistic approach ensures the child is well-rounded and prepared for all aspects of life.

The ninth principle, holistic development, includes five key curricular areas through which Montessori education is delivered. These areas ensure comprehensive coverage of essential skills and knowledge, and are not discrete subjects but interconnected pathways fostering independence, concentration, and curiosity (A. S. Lillard, 2017; Marshall, 2017; Montessori, 2004). The areas are practical life, sensorial, math, language, and cultural studies. However, the specific content, emphasis, and integration of these areas often vary across schools and classrooms, reflecting differences in interpretation, teacher training, and local adaptation of Montessori principles. This variability makes it difficult to aggregate findings across studies, underscoring the need for our meta-analytic and metascientific approach to identify consistent patterns and sources of divergence in the evidence base.

Challenges in Evaluating Montessori Education

Evaluating Montessori education empirically poses a series of conceptual and methodological challenges. The most fundamental is definitional ambiguity. Because the term ("Montessori") is not legally protected or trademarked, implementations range from schools accredited by regional, national, or international Montessori organizations (e.g., the Association Montessori Internationale [AMI] or the American Montessori Society [AMS]) to loosely inspired classrooms (Debs et al., 2022; A. K. Murray & Daoust, 2023). In addition, Montessori education is disproportionately implemented in private settings, with approximately 91% of Montessori schools worldwide being privately funded (Debs et al., 2022). This structural imbalance complicates the

construction of comparable control groups and increases the likelihood of selection effects linked to socioeconomic background. Together, these factors make it difficult to evaluate the method's efficacy on empirical grounds.

Fidelity of implementation, therefore, emerges as both a theoretical and methodological dilemma. Central Montessori features (such as multi-age classrooms, the uninterrupted three-hour work cycle, the presence of complete material sets, and the teacher's guiding rather than instructive role) are often underreported or described inconsistently (A. K. Murray & Daoust, 2023). Such reporting gaps undermine construct validity: effects attributed to "Montessori education" may instead capture local adaptations or even entirely distinct pedagogical logics. At the same time, variability should not be viewed solely as error. As L'Ecuyer (2023) notes, flexibility was integral to Montessori's philosophy.

Practical constraints also hinder experimental rigor. Genuine randomization is virtually impossible because enrollment in Montessori programs usually depends on parental choice or selective admissions. Consequently, quasi-experimental or correlational designs dominate the literature, making causal inference difficult (Debs & Brown, 2017). Limited access to public Montessori settings, ethical concerns about random assignment, and scarce research funding further restrict large-scale trials (Manship, 2023). As a result, evidence often comes from small-sample, cross-sectional studies with limited internal and external validity (Debs et al., 2022).

Measurement poses additional complications. Researchers employ heterogeneous, sometimes bespoke instruments, few of which report reliability or validity indices, creating a mismatch between Montessori's holistic aims and the constructs assessed (Marshall, 2017; A. K. Murray & Daoust, 2023). Sociocultural factors amplify this variability: most studies originate from Western, middle-class contexts, whereas research in racially diverse or Global-South populations remains scarce (Canzoneri-Golden & King, 2023; Debs & Brown, 2017; Moquino et al., 2023). Acknowledging these imbalances is essential for interpreting generalizability.

Finally, researcher allegiance has influenced parts of the literature. A significant proportion of Montessori studies are produced by advocates of the method, which may introduce confirmation or reporting biases (L'Ecuyer, 2023; Marshall, 2017). The limited prevalence of preregistration, open data, or independent replication reinforces the need for metascientific scrutiny. Together, these conceptual, methodological, and contextual limitations explain why conventional meta-analyses have struggled to produce clear conclusions about Montessori's true educational impact (e.g., Demangeon et al., 2023; Randolph et al., 2023).

The Need for a Metascientific Synthesis

Despite a century of experimentation and global presence, the cumulative evidence on Montessori education remains fragmented. Moving the field forward requires not only aggregating results but also examining how methodological rigor, transparency, and fidelity shape those results. This metascientific perspective shifts the question from "Is Montessori effective?" to "How trustworthy is the evidence that claims it is?"

Two recent quantitative syntheses (Demangeon et al., 2023; Randolph et al., 2023) marked an important step toward evidence-based evaluation. However, both studies include several methodological missteps that limited the validity and generalizability of their findings. While both studies attempted to synthesize the impact of Montessori education, they lacked transparency and clarity in their literature search strategies, which undermines reproducibility. Both studies also had much smaller sample sizes, analyzing only 33 and 32 studies, respectively, compared to the 198 studies included in our meta-analysis.

Additionally, both meta-analyses showed methodological weaknesses in the calculation of effect sizes and detection of publication bias. Demangeon et al. (2023) and Randolph et al. (2023) did not sufficiently address biases in small effect sizes or pretest variance in pretest–posttest designs. Furthermore, despite their findings of predominantly nonsignificant results, they interpreted positive effect sizes as meaningful without considering their statistical significance, a practice that calls into question the validity of their conclusions.

The present review addresses this gap by integrating metascientific standards into a large-scale, preregistered meta-analysis. In line with this perspective, the present review was preregistered and adopts an open-science framework aimed at enhancing transparency and reproducibility. Preregistration involves specifying research questions, inclusion criteria, and analytic decisions in advance, thereby reducing researchers' degrees of freedom and minimizing the risk of bias (Lakens et al., 2016; Nosek & Lakens, 2014). Given the absence of prior preregistered syntheses in this domain, the present review was designed as an exploratory analysis without a priori hypotheses. Our approach expands coverage (198 unique studies, >12,000 records), distinguishes design types in computing effect sizes, and models heterogeneity through multilevel and robust-variance estimation frameworks. Publication bias is examined using complementary frequentist and Bayesian techniques (Trim-and-Fill, PET/PEESE, selection models, Mathur–VanderWeele, and RoBMA; Bartoš & Maier, 2020).

Furthermore, beyond methodological appraisal, we introduce a quantitative framework for assessing Montessori fidelity derived from foundational principles (A. S. Lillard, 2017; Montessori, 2004; A. K. Murray & Daoust, 2023). Rather than prescribing authenticity, this analytic model operationalizes fidelity in measurable indicators, enabling systematic exploration of how adherence to Montessori principles relates to outcomes.

Through this synthesis, we aim to address three interrelated questions:

1. What is the magnitude and consistency of Montessori effects across studies?
2. How do methodological rigor and fidelity influence those effects?
3. To what extent do bias and uncertainty shape the current evidential landscape?

By situating these questions within a transparent, preregistered, and openly shared framework, this meta-analysis advances Montessori research from fragmented findings toward a cumulative science that can critically evaluate and refine itself.

Aim and Research Questions

In light of the persistent uncertainty surrounding the educational impact of the Montessori method and the methodological limitations identified in previous syntheses, the present study adopts a metascientific perspective to provide a rigorous and transparent synthesis of the empirical evidence. Our overarching aim is not only to estimate the overall effectiveness of Montessori education on learning outcomes but also to evaluate the credibility of this evidence by examining study quality, fidelity of implementation, and the methodological sources of heterogeneity across studies. Specifically, this meta-analysis addresses the following research questions:

RQ1: What are the main descriptive characteristics of the studies that have examined the Montessori method in educational contexts?

RQ2: What is the average effect of Montessori-based education on learning outcomes across studies?

RQ3: What is the methodological quality of the existing evidence, and how might these quality indicators influence the estimated effects?

RQ4: Which contextual, methodological, or fidelity-related factors moderate the observed heterogeneity in Montessori effects?

Method

Literature Search

This study followed a preregistered systematic search strategy, consistent with PRISMA recommendations (Page et al., 2021, see Appendix B: PRISMA Checklist in the online supplementary file) and APA's Journal Article Reporting Standards (Appelbaum et al., 2018). On January 17, 2023, a project was created in the Open Science Framework (OSF) to host an anonymized preregistration document specifying the methodology and analytical plan for this systematic review and meta-analysis. All materials used in the analyses, including data files and R scripts, are available at <https://osf.io/j9v87/>.

On January 26 and 27, 2023, we conducted a search through the Web of Science, Scopus, PsychINFO, ERIC, and Google Scholar databases using the following search equation in the Topic section: "*Montessori*" AND ("*intervention*" OR "*approach*" OR "*instruction*" OR "*learning*" OR "*method*"). The only restriction imposed on the search was that the papers had to be written in English. To perform the search in Google Scholar and manage the results, we used the Publish or Perish software (Harzing, 2007). For more details on the search results, see Appendix C in the online supplementary file. The final result of this search, after eliminating duplicates, was 2,030 potentially relevant papers. The filtering process was carried out in three phases. The first phase analyzed the papers resulting from the search in the databases, the second analyzed the papers that cite and are cited by the articles included in the first

phase, and the third phase analyzed all the papers cited in the antecedent reviews and meta-analyses. In each phase, the third author screened the papers for potential inclusion by reading the titles and abstracts, using the inclusion criteria outlined below as a reference. This initial screening by a single reviewer was followed by an independent full-text evaluation by two reviewers, ensuring that final inclusion decisions were based on double verification. Papers that passed this initial filter were then independently read in full by the first and third authors. Only those articles that met the inclusion criteria at every stage were included in the final sample. This collection of studies constitutes the sample analyzed in this work. Across all full-text articles assessed for inclusion, the initial inter-rater agreement was 99.93% (Cohen's Kappa = .984). Any disagreements were discussed until 100% consensus was reached.

Figure 1 provides a detailed PRISMA flowchart of the entire search process, showing that the final number of unique papers included was 198.

Selection Criteria

All studies considered for potential inclusion were required to meet the following inclusion criteria:

C1: Empirical studies evaluating the impact of interventions based on the Montessori method.

C2: Conducted in educational contexts (at any academic level).

C3: Evaluating impact through academic performance and other variables related to learning in educational settings.

C4: Written in English.

Studies were restricted to those published in English. This decision was made to ensure consistency and reliability in the screening and coding process across reviewers. Given that the databases consulted primarily index peer-reviewed literature in English, we expect that the majority of studies with international visibility were captured. However, we acknowledge that relevant studies published in other languages may have been excluded. To enhance transparency, Appendix D provides detailed examples illustrating the application of these criteria, including representative cases of excluded studies that did not meet one or more inclusion criteria, and one fully eligible study that met all requirements.

Data Extraction and Coding

The final sample of articles included in the analysis was 198. We coded the general descriptive characteristics of these studies, including authors and year of publication, type of study, sample type, total sample size (and sample size for each group), age of the participants, proportion of men/women, educational level, person conducting the intervention, whether the intervention was implemented as part of the study, duration

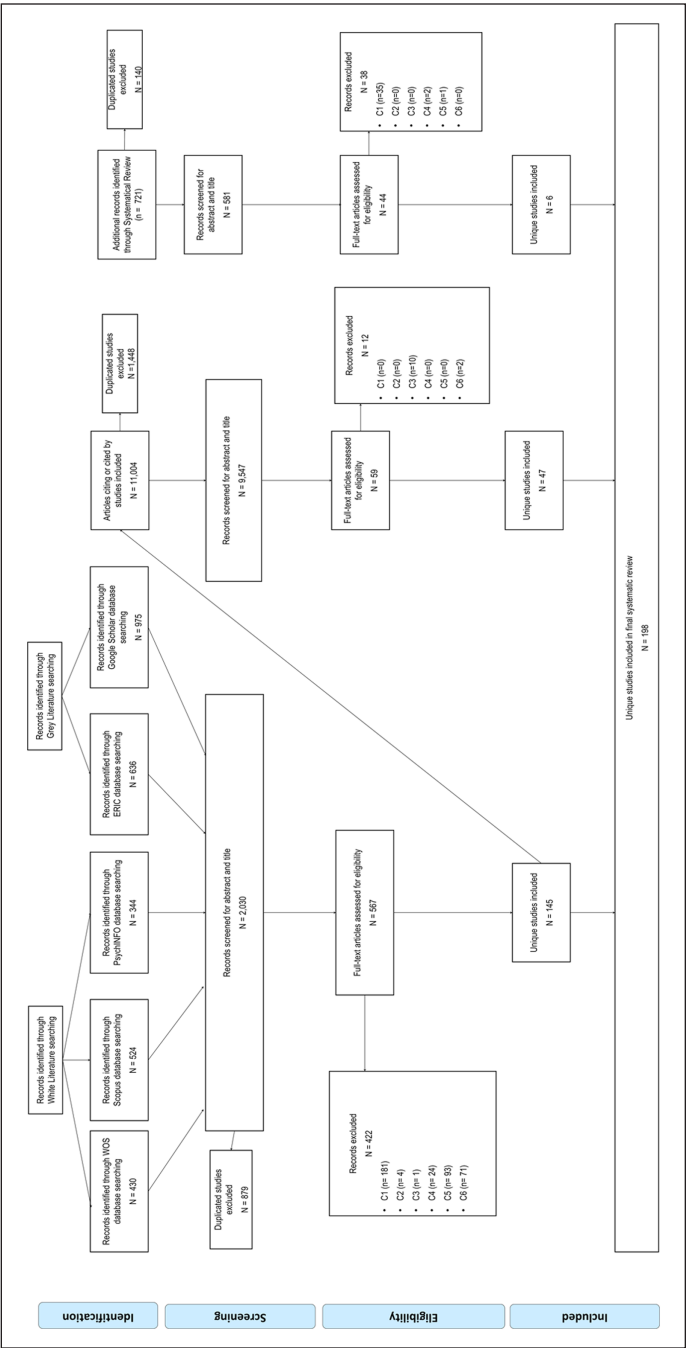


Figure I. PRISMA flowchart.

of the intervention, dependent variables analyzed, and the reliability of the measurement tools used to assess the dependent variables. See Table E1 of Appendix E in the online supplementary file for more detailed information. A concise summary of the main descriptive characteristics (e.g., study type, region, educational level, and intervention status) is also presented in Table E1.1. (Appendix E).

To address concerns raised by previous researchers who attempted to synthesize studies of Montessori effects, we also coded the methodological characteristics used in the included studies. To this end, we employed the scale created by Ferrero et al. (2021a) to analyze the methodological quality of intervention programs in educational contexts. This scale comprises 17 items designed to assess adherence to characteristics that directly impact the quality and methodological rigor of educational interventions. These characteristics include elements such as preregistration of studies, random assignment of participants or conditions, blinding of relevant stakeholders, and analysis of potential pre-intervention differences. Each item is rated on a scale: (a) positive, when the study met the criterion; (b) negative, when the study did not meet the criterion; and (c) unknown, when the study did not provide sufficient information. In the original scale, item 10 analyzed the fidelity of the intervention (in that case based on multiple intelligences; see Ferrero et al. [2021a]). In our study, we replaced this item with a new table that analyzes the nine foundational principles proposed by Maria Montessori, which underpin any intervention based on the Montessori method (Montessori, 2004). Since the ninth principle outlines five curricular areas of intervention within the Montessori method, we evaluated the implementation of these areas in each study. For more details on the principles, curricular areas, and their operationalization in the assessment process, see Appendix A in the online supplementary file. Beyond labeling studies as pre–post or post-only, we coded assignment and risk-of-bias features adapted from Ferrero et al. (2021a). Specifically, we recorded whether studies reported: randomization of participants and/or groups; blinding of participants, implementers/teachers, and analysts; baseline equivalence (socioeconomic variables); analysis of pretest scores; presence of an active control group; implementer training and fidelity checks; sufficient information to replicate the intervention and the outcome measures; reporting of at least one key comparison; and open data access (see Appendix G). We also distinguished studies that implemented a Montessori intervention during the study from those comparing cohorts already enrolled in Montessori schools (Appendix E, column “Intervention”: Yes vs. No). These item-level codings were used for descriptive summaries and as candidate moderators in Appendix L. For transparency, Appendix L begins with a brief guide that lists the reference categories and contrasts used in all subgroup and meta-regression tables.

Computation of Effect Sizes and Statistical Analyses

In addition to the qualitative descriptive information, we also coded the information needed to estimate the effect size of the intervention. Four designs emerge from the papers included in the review. Firstly, we have designs with pretest and posttest measures that include a control group (hereinafter pretest–posttest); secondly, there are

designs with a control group where there is only a posttest measure after the intervention (hereinafter posttest-only); thirdly, we found single group studies with pretest and posttest (single group pretest–posttest); and fourthly, we found studies in which only one group was assessed after the intervention. For our work, the first two will be of special relevance in the meta-analysis. For the first group of studies (pretest–posttest), we estimated the effect size using the standardized mean change difference score ($\bar{g}_\Delta$) recommended by Morris (2008). We calculated the effect size using Eqs. 8–10, and the variance of the effects with Eq. 25 (Morris, 2008). For the calculation of this variance, estimating the correlation between pretest and posttest scores is essential. However, since this information is often unavailable, we followed Rosenthal's (1991) recommendations and assumed a correlation of .70.

In addition, many of the included papers only reported test results after the intervention (post-only designs). For these papers, as well as for the papers in the pretest–posttest designs, we calculated a standardized mean difference of posttest scores ($\bar{g}_p$). This effect size was calculated following the recommendations of Borenstein et al. (2009), applying Eqs. 4.18 and 4.19, as well as including the correction factor j estimated with Eqs. 4.22 and 4.23. We estimated the variance for these effects with Eqs. 4.20 and 4.24. Note that while the $\bar{g}_\Delta$ effect could only be calculated with the pretest–posttest designs, the $\bar{g}_p$ effect could be calculated for the pretest–posttest ($\bar{g}_{p-pp}$) and posttest-only ($\bar{g}_{p-p}$) designs. In the case of the single group pretest–posttest design, we estimated the standardized mean change using Becker's d (Becker, 1988) by applying Eq. 2 and Eq. 13 for the variance. Finally, studies without a control group or pretest were not considered for the meta-analysis (for more details on the formulas used, see Appendix F in the online supplementary file). Positive values of the effect sizes represent greater benefits in favor of the Montessori group, and negative values index the contrary. We multiplied by -1 those effects in which it was necessary to keep the mentioned direction (e.g., number of errors). For studies with both pre- and posttest measures, we computed two parallel estimates to ensure comparability across metrics: (a) $\bar{g}_\Delta$, the standardized mean change using pretest SDs (Morris, 2008), and (b) $\bar{g}_{p-pp}$, the posttest standardized mean difference using posttest SDs (Borenstein et al., 2009). This dual computation allowed us to assess whether differences in the standardization metric (pre- vs posttest SD) contributed to heterogeneity or inflated effect sizes. All effect sizes ($\bar{g}_\Delta$, $\bar{g}_p$, and d) represent standardized mean differences, corrected for small-sample bias when applicable, and can therefore be interpreted on the same scale as Hedges' g . Conventionally, values around 0.2, 0.5, and 0.8 are described as small, medium, and large effects (Cohen, 1988), although effect magnitudes in education research tend to be smaller on average (Kraft, 2020). Subsequent analyses examined these metrics separately and as potential moderators.

A large number of the included papers contributed more than one effect size for the same sample, which may compromise the independence of the results. It is common in conventional meta-analyses to assume independence among effect sizes. In an attempt to address this potential methodological shortcoming, the robust variance estimation approach has been developed (RVE; Hedges et al., 2010). This method estimates the

correlation matrix and sets the weights according to a correlated or a hierarchical structure. Different simulation studies (Hedges et al., 2010) have shown that RVE has very good precision in estimating both effect sizes and confidence intervals, even with small study samples ($m = 10$), having a large number of dependent estimates per study ($k = 10$). We used the *robumeta* package for R (Fisher et al., 2017) for the implementation of RVE conducted in the main analyses. We chose a correlated dependence model with small-sample corrections (Tipton, 2015).

In the first analytical step, we performed calculations of the overall total effect of the Montessori-based intervention. Due to the characteristics of the included designs as well as the calculation of the effect sizes associated with each design, our work calculated an overall effect for each one $\bar{g}_\Delta$, $\bar{g}_{p-pp}$, $\bar{g}_{p-p}$ and d_{single} . The usual heterogeneity indexes, τ^2 and I^2 , were computed. To identify studies with outlying outcomes, we fitted a multilevel model with the *rma.mv()* function of *metafor* (Viechtbauer, 2010) and estimated the studentized residuals (>2) and Cook's distance ($>4/n$). For studies with a pretest measure, we estimated an RVE model with the scores obtained in the pretest measure ($\bar{g}_{pre}$), considering this effect as the difference in the baseline between groups.

Finally, we examined the potential influence of various variables coded as possible moderators of the effect. First, we analyzed the descriptive variables of the studies as potential moderators. These included study design, type of dependent variables, baseline differences, year of publication, student age (in years), type of intervention (with or only test), intervening agent, type of work, educational stage, sample type, and continent (For more details on how each moderator is operationalized and its different levels, see Appendix L).

Second, we explored the potential moderating role of factors related to methodological quality (Ferrero et al., 2021a) and the principles and areas of intervention proposed as fundamental in Montessori-based interventions (Montessori, 2004). In line with critiques of using quality scores (Jüni et al., 1999), we refrained from aggregating the items of the scale into a single score. When analyzing the moderating factors, we analyzed whether the different levels differed from each other, for which we used the Wald test using the *clubSandwich* package (J. Pustejovsky, 2024). This test allows you to examine the contrasts of the regression model calculated using a cluster of estimates for the variance-covariance matrix, as well as a small-sample correction for the *p-value* (Tipton & Pustejovsky, 2015).

Studies reporting null or contradictory findings are less likely to appear in print, a phenomenon commonly referred to as publication bias. Various methods were available to detect potential publication biases in meta-analyses (Carter et al., 2019). To assess possible publication bias in our study, we employed the methodology outlined by Román-Caballero et al. (2022).

We used four techniques to detect publication biases and adjust the estimated average effects. Two of these techniques analyzed aggregate data, and two analyzed multilevel data. This multimethod approach, which integrated different detection procedures, was demonstrated to be a robust strategy (Carter et al., 2019). The

techniques included: (i) the trim-and-fill method (with L0 and R0 estimators) and the selection model for aggregate data, (ii) the RVE regression-based approaches (RVE PET and RVE PEESE), and (iii) Mathur and VanderWeele's sensitivity analysis. For further details on these methods, see Appendix M in the online supplementary file.

To aggregate the intra-study data, we used the package *MAAd* (Del Re & Hoyt, 2014). For the model detection test, we will use the package *weightr* (Coburn & Vevea, 2019), and we will use *PublicationBias* (Mathur & VanderWeele, 2020) for Mathur and VanderWeele's sensitivity analysis. In the RVE Meta-regression testing approaches, we used a modified form of the sample variance, as well as a variance stabilizing transformation for the standardized mean contrast to avoid artifactual dependence between the effect size and its precision estimate (J. E. Pustejovsky & Rodgers, 2019; see Appendix M).

Finally, due to the anomalies encountered with the previous methods for detecting publication bias and the limitations of frequentist inferences on nonsignificant results, we conducted a robust Bayesian meta-analysis with the aggregated data (Bartoš et al., 2023). This approach enabled us to estimate 36 meta-analytic models, considering the presence or absence of effect, heterogeneity, and publication bias. Additionally, it allowed us to estimate the average Bayesian model for effect size, effect heterogeneity, and the possible presence of publication bias. For this estimation, we utilized the R package *RoBMA* (Bartoš & Maier, 2020). All data and the R script used for the analyses are fully available at <https://osf.io/j9v87/>.

Results

Descriptive Analyses

Figure 2 illustrates the number of works included over the years indicating a marked increase in research activity on the Montessori method, particularly since the early 2000s. The data from the 198 papers included in the study allowed us to analyze a total sample of 40,758 students.

Most of the studies were carried out with students in preschool (29.9%) and primary school (44.3%). A much smaller proportion focused on secondary school (5.0%), while 7.0% were classified as other educational levels and 13.9% did not specify the level. The reported age range of the students was between 2.3 and 17.5 years. Of the included studies, 82.8% used a normal study population, 11.4% used students with functional diversity, and 5.8% used both. On average, of the studies reporting the gender ratio of the sample, 49.1% were male students. Of the total studies, there were 133 articles, 62 doctoral dissertations, one book, one book chapter, one conference proceeding, and one action research project. In all, 53.2% of the studies were conducted in North America, 18.8% in Asia, 14.3% in Europe, 7.8% in Oceania, 4.5% in Africa, and 1.3% in South America. From the information in the articles, we observed that 35.8% of the papers providing data to analyze the effect of the Montessori method used a pretest–posttest design with a control group, 48.3% used a posttest-only design with a control group, 5.8% used a pretest–posttest single-group design (a single

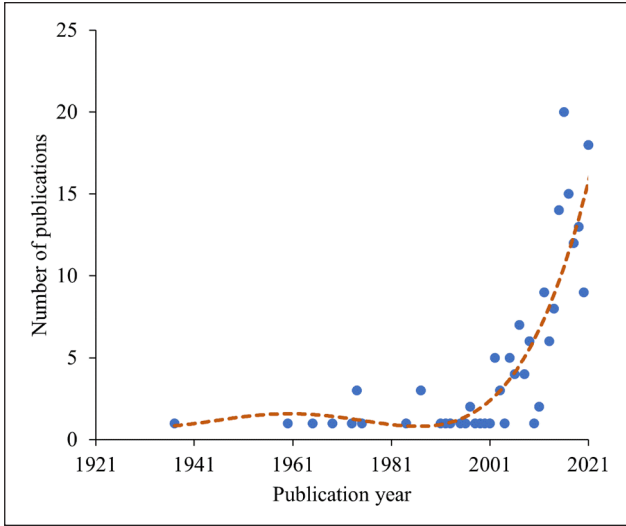


Figure 2. Number of publications included in the review by years. The red line represents the developing trend of publications across years.

Montessori group), and 10.0% unfortunately used a posttest measure without a control group. From the information contained in the papers, we could see that in 63.7% of the studies the person who carried out the intervention was the researcher, in 10% the teacher, in 16.5% both, and in 9.7% of the papers no information was reported. Of the total number of studies analyzed, in 52% of them, for the study, an intervention based on the Montessori method was carried out, while in 41.5% they analyzed the data collected without proposing a controlled intervention. That is, they measured a specific variable in schools where teaching was already being carried out through the Montessori method. In 6.3% of the studies no information was provided on this aspect. Of all the variables analyzed, 70.7% were academic achievement, 11.3% cognitive abilities, 1.3% creativity, 7% emotions, 3.4% motor skills, and 6.3% social skills.

Regarding the methodological quality of the studies analyzed, Figure 3 summarizes the details of the quality of the 198 articles. As can be seen, a large number of studies did not provide sufficient information to be able to assess the quality factors. To highlight some examples, the items related to the blinding of participants, instructors, or analysts (Items 4, 5, and 6) could not be evaluated in 93.63%, 75.49%, and 87.25%, respectively, and using tests to measure the variable depending on satisfactory validity (Item 15) could not be evaluated in 76.47% of the studies. In Figure 3, we also observe that a substantial number of studies did not meet the analyzed criteria. Notably, preregistration of studies (Item 1) was absent in 99.02% of cases, the use of an active control group (Item 11) was not met by 78.43%, providing sufficient information to replicate the intervention (Item 12) was lacking in 81.37%, and offering open data (Item 17) was not met by 97.55%. Conversely, a relatively high percentage of studies

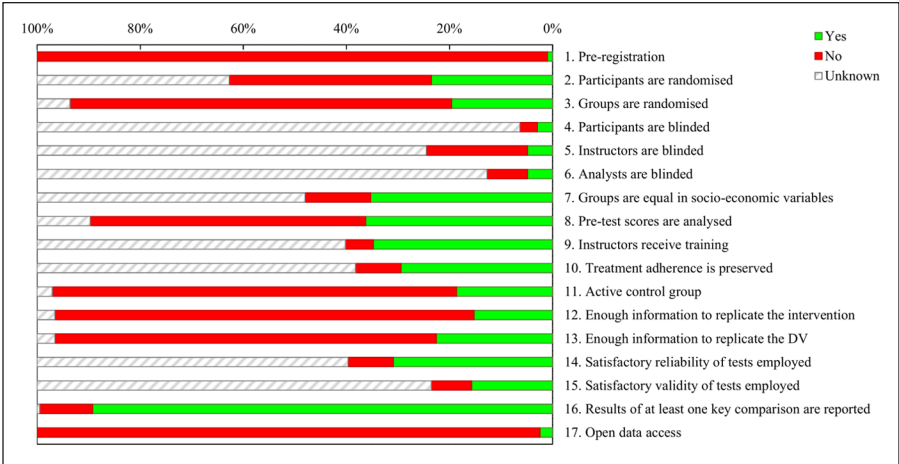


Figure 3. Graphical representation of the results after evaluation of the methodological quality of the included studies.

Note. “Unknown” values are represented using patterned shading to ensure visibility in grayscale.

satisfactorily fulfilled the criterion of analyzing at least one key variable (Item 16), with 89.22% of studies meeting this requirement. For a detailed overview of the quality item values assessed for each study, please refer to Appendix G in the online supplementary file. An aggregated overview of compliance with the 17 methodological quality items is provided in Table G1.1 (Appendix G), which reports the frequency and percentage of studies meeting each criterion.

Regarding the principles and curricular areas, Figure 4 presents a detailed summary. Panel A displays the evaluation of the principles, while Panel B focuses on the curricular areas. For the principles evaluated, most studies lacked sufficient information to assess compliance, with the exception of the item related to general compliance with the curricular areas (Item 9), which demonstrated a high compliance rate of 73.53%. The percentages of studies that did not provide information for evaluating the principles ranged from 60.29% to 64.71%.

In terms of the curricular areas, the most frequently addressed areas in the interventions were language and mathematics, appearing in 35.78% and 28.92% of the studies, respectively. For a detailed overview of the coverage of Montessori principles as well as curricular areas in each of the included studies, please refer to Appendix H in the online supplementary file. A summary of these results is presented in Table H1.1 (Appendix H), showing the proportion of studies addressing each Montessori principle and curricular area.

Quantitative Analyses

Overall Effect. Of the 198 studies, only 114 presented quantitative information that could be analyzed in our meta-analysis. Of these, 42 papers had a pretest–posttest

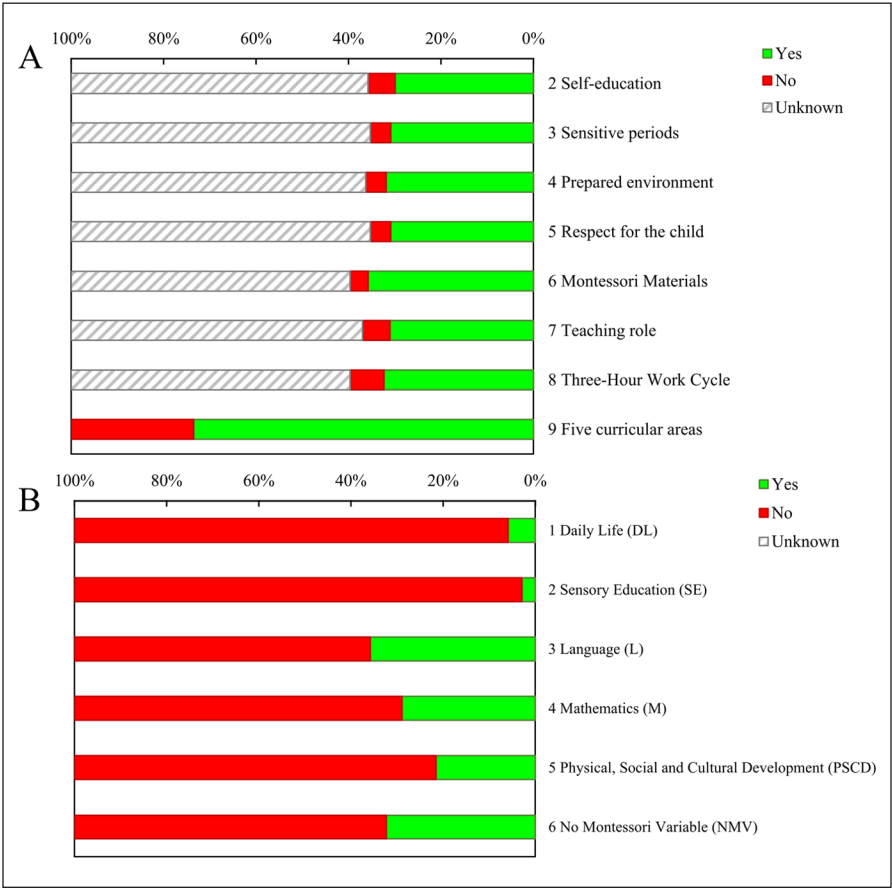


Figure 4. Graphical representation of the results after evaluating compliance with the principles of fidelity with the Montessori method (Panel A) and with the curricular areas (Panel B).

Note. The principle “absorbent mind” was conceptually identified but not included in the figure or analyses because it could not be operationalized at the study-report level. Unknown values are represented using patterned shading to ensure visibility in grayscale.

design, 53 had a posttest-only design, six had a pretest–posttest single group design, and 12 were a posttest study without control.

The overall effect of using the Montessori method for the standardized mean change of the pretest–posttest effect resulted in a significant positive effect, $\bar{g}_\Delta = 1.03$, 95%CI [0.66, 1.40], $p < .001$, with heterogeneity of $\tau^2 = 0.75$; $I^2 = 95.80\%$ ($M = 32$, $k = 149$). In the case of the effect measured through the standardized mean difference in the post-test, the overall total effect (considering the effects from the pretest–posttest and post-test-only designs) was $\bar{g}_p = 0.84$, 95%CI [0.53, 1.14], $p < .001$, with a heterogeneity of

$\tau^2=2.53$; $I^2=98.66\%$ ($M=81$, $k=395$). A closer analysis showed that the posttest effect by isolating the pretest–posttest design was $\bar{g}_{p-pp}=1.01$, 95%CI [0.62, 1.39], $p<.001$, with a heterogeneity of $\tau^2=1.11$; $I^2=95.74\%$ ($M=35$, $k=155$), while isolating the posttest-only design was $\bar{g}_{p-p}=0.70$, 95%CI [0.27, 1.13], $p<.001$, with a heterogeneity of $\tau^2=3.40$; $I^2=98.98\%$ ($M=48$, $k=240$).

Next, we analyzed the presence of outliers in the effect sizes. This analysis revealed a significant number of outliers (see Appendix I in the online supplementary file for details), which were excluded from subsequent analyses. Despite the removal of these outliers, the average effect for pretest–posttest designs remained positive and significant, with an effect size of $\bar{g}_\Delta=0.74$, 95%CI [0.40, 1.08], $p<.001$, with a heterogeneity of $\tau^2=0.42$; $I^2=93.56\%$ ($M=27$, $k=135$), as well as the total posttest effect of $\bar{g}_p=0.80$, 95%CI [0.56, 1.04], $p<.001$, with a heterogeneity of $\tau^2=1.64$; $I^2=98.00\%$ ($M=75$, $k=364$). The posttest effect isolating the pretest–posttest designs was $\bar{g}_{p-pp}=0.97$, 95%CI [0.62, 1.31], $p<.001$, with a heterogeneity of $\tau^2=0.81$; $I^2=95.51\%$ ($M=34$, $k=142$), and isolating the posttest-only design was $\bar{g}_{p-p}=0.65$, 95%CI [0.33, 0.97], $p<.001$, with a heterogeneity of $\tau^2=1.91$; $I^2=98.26\%$ ($M=43$, $k=222$). Forest plot representations of these effects are provided in Appendix J in the online supplementary file (Figures J1 to J4).

Finally, the average effect for studies with a single group pretest–posttest design was positive but not significant, $d_{single}=0.64$, 95%CI [−0.29, 1.56], $p=.118$, with a heterogeneity of $\tau^2=0.35$; $I^2=90.86\%$ ($M=4$, $k=7$). These studies will be excluded from further analysis due to the limitations in their interpretation arising from the lack of appropriate control measures.

Analysis of Baseline Differences Between Groups. For studies that included a pre-intervention measure, we were able to analyze whether there were systematic differences between the Montessori and control groups in the variables measured prior to the intervention. To find this out, we performed a meta-analysis of the measures $\bar{g}_{pre}$. The overall average effect for the pretest-post studies was not significant, $\bar{g}_{pre}=0.11$, 95%CI [−0.01, 0.23], $p=.077$, with a heterogeneity of $\tau^2=0.10$; $I^2=71.18\%$ ($M=27$, $k=135$); see Appendix K, Figure K1 in the online supplementary file.

Table K1 (Appendix K, in the online supplementary file) presents the analysis of baseline differences across the various dependent variables. Although none of the variables showed a statistically significant difference, both academic achievement and motor skills approached significance. This suggests that while there were no definitive pre-intervention differences between the Montessori and control groups, certain areas like academic and motor skills may warrant further investigation.

Moderator Analyses. To explore potential moderators, we began by analyzing potential differences between the two designs (pretest–posttest vs. posttest-only). This factor did not appear to significantly modulate the heterogeneity of the effect (see Appendix L, Table L1 in the online supplementary file). Detailed results of the moderation analyses are presented in Appendix L. For transparency, each block of moderators is reported

with both individual (subgroup) contrasts and a complementary multivariate meta-regression table specifying the included moderators, reference categories, and model-fit indices. Therefore, moving forward, we analyzed the effects of the moderating variables collectively. In this combined analysis, all effects were collapsed into a single effect size $\bar{g}$ where the pretest–posttest design corresponded to $\bar{g}_\Delta$ and the posttest-only design corresponded to $\bar{g}_p$.

We first analyzed potential moderating variables related to the descriptive characteristics of the studies. The isolated analysis revealed that only the sample type variable had a significant moderating effect, $F(2, 3.7) = 8.59, p = .040$, where the effect was significant for the typically developing students ($\bar{g} = 0.56 [0.31, 0.80]$) but not significant for samples that included students with disabilities or combined populations. We also examined the type of outcome variable (academic achievement, cognitive abilities, social skills, emotional measures, creativity, and motor skills) as a categorical moderator (see Appendix L, Table L1). Differences among these outcome types were not statistically significant ($F = 0.32, p = .889$). Within the academic achievement category, which included diverse indicators such as standardized tests, teacher-assigned grades, and researcher-developed tasks, considerable variability may remain. This residual variability may partly account for the high heterogeneity observed in this domain and highlights the need for future studies to differentiate academic outcomes by assessment type and domain, as well as to further explore Montessori's curricular areas as distinct learning domains.

Recently, it has been proposed that a more appropriate way to capture the educational impact of Montessori pedagogy may lie in the variability of students' achievement scores rather than in their mean levels, as greater dispersion could reflect more individualized learning growth (Scippo, 2024). To examine this hypothesis (raised during the editorial review process), we conducted a non-preregistered exploratory analysis focusing on outcome dispersion as an indicator of learning diversity. Using the same dataset as the main analyses, we computed the log variability ratio (lnVR) and the log coefficient of variation ratio (lnCVR) to compare within-group variability between Montessori and control conditions.

Results indicated slightly lower dispersion in Montessori groups at posttest (lnVR = $-0.10, p = .036$; lnCVR = $-0.26, p < .001$), no baseline differences, and no significant association between variability and mean effect sizes. In additional exploratory analyses, we further examined whether outcome dispersion differed as a function of the breadth of explicitly reported Montessori principles (e.g., the number of core Montessori principles explicitly described as present in each study, rather than a direct measure of implementation fidelity). These analyses did not indicate greater variability in studies reporting a broader set of Montessori principles (see Appendix N, Table N2).

Taken together, although exploratory and not preregistered, these findings suggest that greater outcome variability is not a consistent empirical feature of Montessori education as currently implemented and reported in the literature. However, they do not constitute a definitive test of whether increased variability may emerge under conditions of rigorously assessed high-fidelity implementation, which remains insufficiently and inconsistently documented in primary studies (see Appendix N).

To evaluate whether multiple moderators jointly explained the observed heterogeneity, we next fitted a multilevel meta-regression model including sample type, type of work, and intervenor as predictors (see Appendix L, Tables L1 and L1.1). The omnibus test for this specification was significant, $F(5, 58)=4.47$, $p=.002$ (cluster-robust SEs). However, the inclusion of these moderators led to only a modest reduction in residual between-study heterogeneity, from $\tau^2=1.051$ in the null model to $\tau^2=0.913$ ($\approx 13\%$ reduction). This indicates that, although certain moderators, particularly the role of the intervenor and the sample type, were statistically significant, substantial between-study variance remains unexplained.

Regarding methodological characteristics (Appendix L, Table L2), individual contrasts did not yield consistent moderation effects. To assess their joint contribution, we fitted a multilevel meta-regression including training/experience of the instructor, fidelity of intervention, active control group, replicate intervention, reliability of the DV, and key comparison (Appendix L, Table L2.1). The omnibus test was significant, $F(10, 53)=4.48$, $p=.0001$ (cluster-robust SEs). However, residual heterogeneity decreased only modestly, from $\tau^2=1.051$ (null) to $\tau^2=0.881$ ($\approx 16.2\%$ reduction), indicating that methodological features account for a limited share of between-study variance. We note that “NA” indicates insufficient reporting in the primary studies (not a substantive category), and “Open Data” was present in only one study.

Finally, we examined Montessori’s principles and curricular areas as potential moderators. Consistent with the univariate contrasts (Appendix L, Table L3), we fitted a multilevel meta-regression including prepared environment and five curricular areas (Appendix L, Table L3.1). The omnibus test was significant, $F(3, 60)=5.02$, $p=.0036$ (cluster-robust SEs). Residual heterogeneity decreased from $\tau^2=1.051$ (null) to $\tau^2=0.887$ ($\approx 15.6\%$ reduction). Notably, studies not meeting the prepared environment criterion showed larger effects relative to “NA” ($\beta=1.237$, $p=.0018$), whereas the contrast for “Yes” vs “NA” was not significant ($\beta=0.307$, $p=.173$). The five curricular areas contrast showed a negative trend ($\beta=-0.495$, $p=.053$), suggesting smaller effects when curricular-area alignment was explicitly reported. As in prior blocks, “NA” indicates insufficient reporting in primary studies and should not be interpreted as a substantive level.

Together, these models (Appendix L, Tables L1.1–L3.1) illustrate how different sets of moderators incrementally reduce residual heterogeneity, clarifying the sources of between-study variability.

Quality-Stratified Sensitivity Analysis. To examine whether methodological rigor might influence the estimated effects, we conducted an additional meta-analysis stratified by study quality (see Appendix L, Table L4). Studies were classified into three tiers according to the presence of core quality features: randomization or active control, baseline handling, measurement reliability/validity, and implementation or replication information. No study met all Tier-A requirements, indicating a complete absence of fully rigorous designs in this literature. In the overall dataset, pooled effects were $\bar{g}=0.58$ (95% CI [0.31, 0.84]; $k=224$, $m=47$) for Tier B and $\bar{g}=0.59$ (95% CI [0.13, 1.05]; $k=130$, $m=23$) for Tier C. The Tier B–C contrast was nonsignificant ($p=.26$, CR2). Restricting the analysis to outcomes with documented reliability or validity

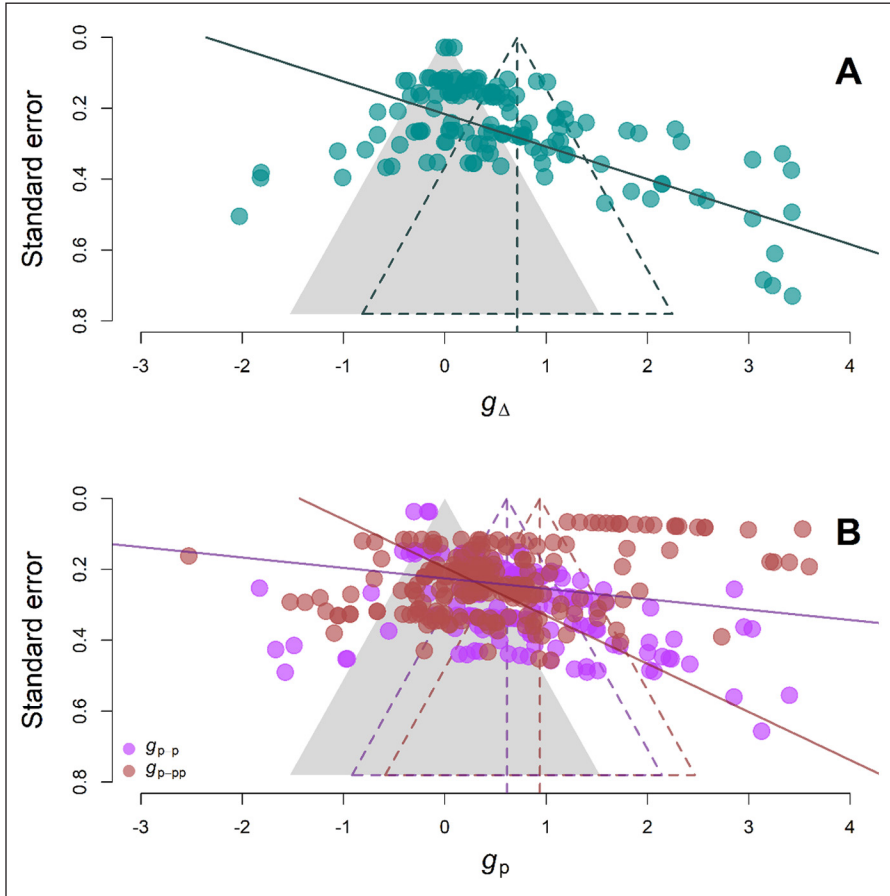


Figure 5. Funnel plot for the meta-analysis of pretest–posttest effect (Panel A) and posttest-only effect (Panel B). The legend in Panel B represents the distinction between the effects from the two designs pretest–posttest (g_{p-pp}) and posttest-only (g_{p-p})

yielded $\bar{g} = 0.46$ (95% CI [0.06, 0.85]; $k = 80$, $m = 18$) for Tier B and $\bar{g} = 1.36$ (95% CI [−3.75, 6.47]; $k = 16$, $m = 2$) for Tier C, again with a nonsignificant difference ($p = .26$). Because the degrees of freedom for Tier C were extremely low ($df = 1$), these results should be interpreted cautiously. Overall, the lack of Tier-A studies and the similarity between Tiers B and C suggest that the large effects reported in the literature are not clearly attributable to methodological rigor but rather to persistent limitations in study design and reporting quality.

Publication Bias. Figure 5 represents the distributions of the effect sizes analyzed in this work through funnel plots. Panel A (top) represents the pretest–posttest effects ($\bar{g}_\Delta$)

while Panel B (bottom) represents the posttest effects, both for pretest–posttest (effect $\bar{g}_{p-pp}$) and posttest-only designs (effect $\bar{g}_{p-p}$). The solid line in each graph represents the linear regression (Egger's test) that best fits all effects for each effect mode. A visual inspection of the dispersion of the effects suggests some asymmetry in the results, especially in the case of the pretest–posttest effects.

We applied four complementary methods to detect and adjust for potential publication bias (trim-and-fill, PET/PEESE, selection models, and sensitivity analysis). Across these approaches, results were mixed regarding the presence of bias, although corrected estimates differed substantially from the unadjusted effects. For further details on these analyses, please refer to Appendix M in the online supplementary file.

Consistent with our preregistered multimethod plan, and assuming the limitations that frequentist analyses have in trying to analyze nonsignificant results, we performed a robust Bayesian meta-analysis for each effect size (Ciria et al., 2023). Thus, a robust Bayesian meta-analysis for the pretest–posttest design found weak evidence against the effect, $BF_{10}=0.71$, with a mean model-averaged estimate $\bar{g}_{\Delta}=0.13$, 95% CI $[-0.08, 0.69]$ (see Figure 6), and strong evidence in favor of the heterogeneity, $BF^{rf}=2,67E+42$, with mean model-averaged estimate $\tau=0.67$, 95% CI $[0.49, 0.94]$. This analysis yielded strong evidence consistent with publication bias, $BF_{pb}=31.90$. In the case of the posttest effect (both pretest–posttest and posttest-only design), robust Bayesian meta-analysis found weak evidence against the effect, $BF_{10}=0.700$, with mean model-averaged estimate $\bar{g}_p=0.15$, 95% CI $[-0.12, 0.75]$ (see Figure 7), strong evidence in favor of the heterogeneity, $BF^{rf}=\infty$, with mean model-averaged estimate $\tau=0.99$, 95% CI $[0.76, 1.23]$, and strong evidence in favor of the publication bias, $BF_{pb}=14.85$. In order to further investigate this effect, we performed the analysis individually for each design. Thus, the result for the posttest effect for pre–post designs found weak evidence against the effect, $BF_{10}=0.94$, with mean model-averaged estimate $\bar{g}_{p-pp}=0.21$, 95% CI $[-0.18, 0.95]$ (see Figure 8), strong evidence in favor of the heterogeneity, $BF^{rf}=2.4753e+48$, with mean model-averaged estimate $\tau=0.74$, 95% CI $[0.55, 1.06]$, and strong evidence in favor of the publication bias, $BF_{pb}=12.76$. In the case of posttest-only design, we found weak evidence against the effect, $BF_{10}=0.44$, with mean model-averaged estimate $\bar{g}_{p-p}=0.04$, 95% CI $[-0.44, 0.63]$ (see Figure 9), and strong evidence in favor of the heterogeneity, $BF^{rf}=\infty$, with mean model-averaged estimate $\tau=1.00$, 95% CI $[0.78, 1.26]$, and strong evidence in favor of the publication bias, $BF_{pb}=15.76$. RoBMA averages across 36 submodels that variably include or exclude a nonzero effect, random-effects heterogeneity, and selection adjustment. Posterior mass concentrated on heterogeneous, selection-adjusted models (including weight-function models $[\omega]$, PET, and PEESE), consistent with a publication-bias pattern: posterior inclusion probabilities approached 1.00 for heterogeneity and 0.92–0.99 for publication bias, whereas evidence for a nonzero, grand-mean effect was modest (≈ 0.20 – 0.40). Specifically, estimates were: pretest–posttest with control (heterogeneity ≈ 1.00 ; bias ≈ 0.97 ; effect ≈ 0.41), combined posttest across designs (≈ 1.00 ; 0.99 ; 0.21), posttest within pretest–posttest designs (≈ 1.00 ; 0.92 ; 0.39), and posttest within posttest-only designs (≈ 1.00 ; 0.96 ; 0.27). These outcomes indicate strong evidence for bias and heterogeneity but only modest evidence for a genuine

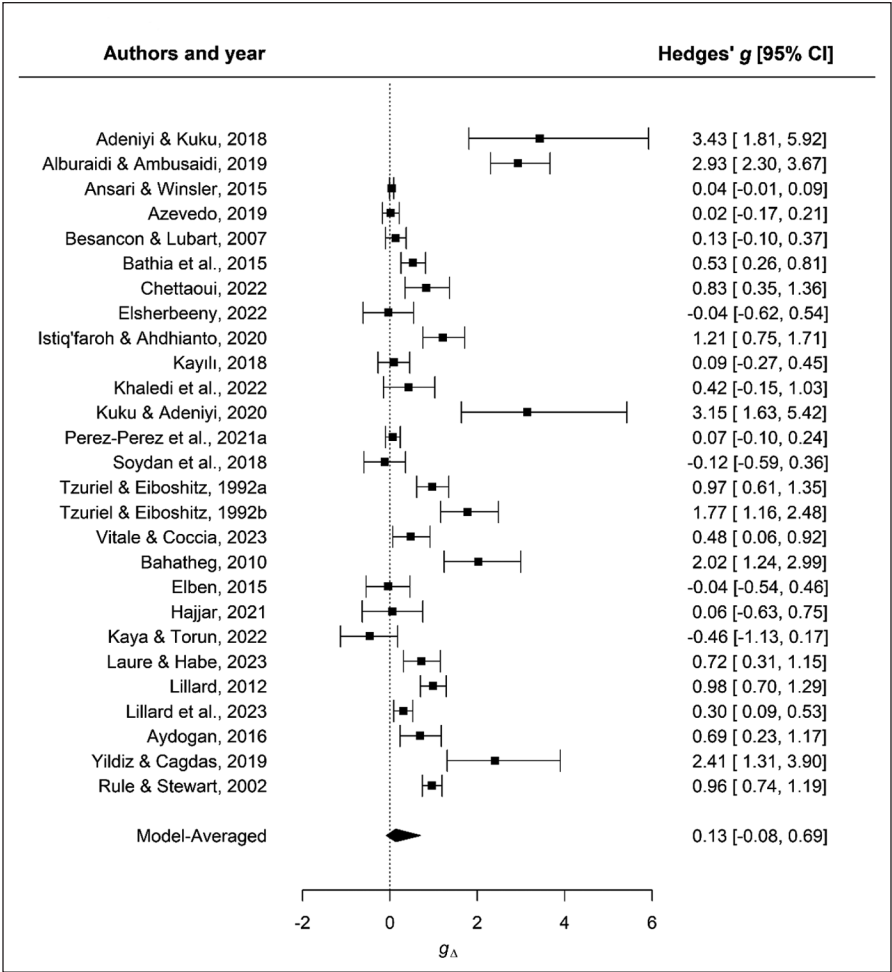


Figure 6. Forest plot for the robust Bayesian meta-analysis of pretest–posttest effect ($\bar{g}_A$).

effect. Full posterior inclusion probabilities and the five submodels with the highest weights are reported in Appendix M (Tables M6–M9).

Discussion

The primary goal of this preregistered systematic review was to rigorously, transparently, and comprehensively analyze the available evidence on the educational impact of the Montessori method. Despite the growing interest in Montessori education, there remains limited consensus regarding its overall effectiveness and applicability across diverse educational contexts and student populations.

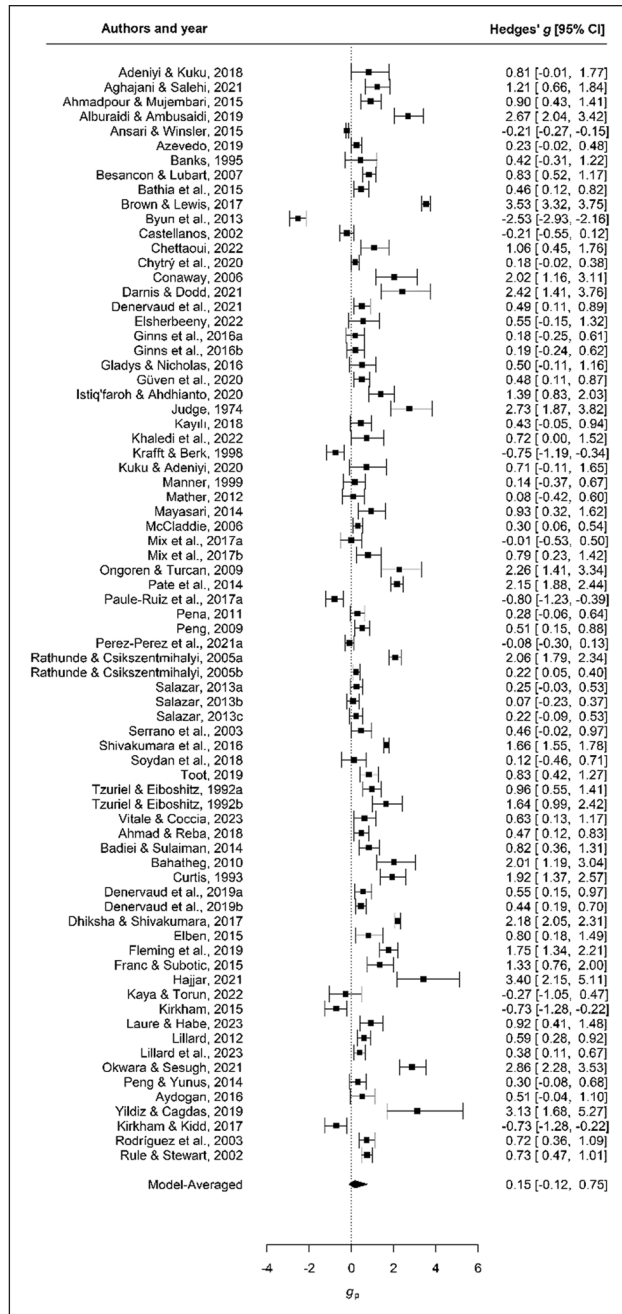


Figure 7. Forest plot for the robust Bayesian meta-analysis of only-post effect for the pretest-posttest and posttest-only design ($\bar{g}_p$).

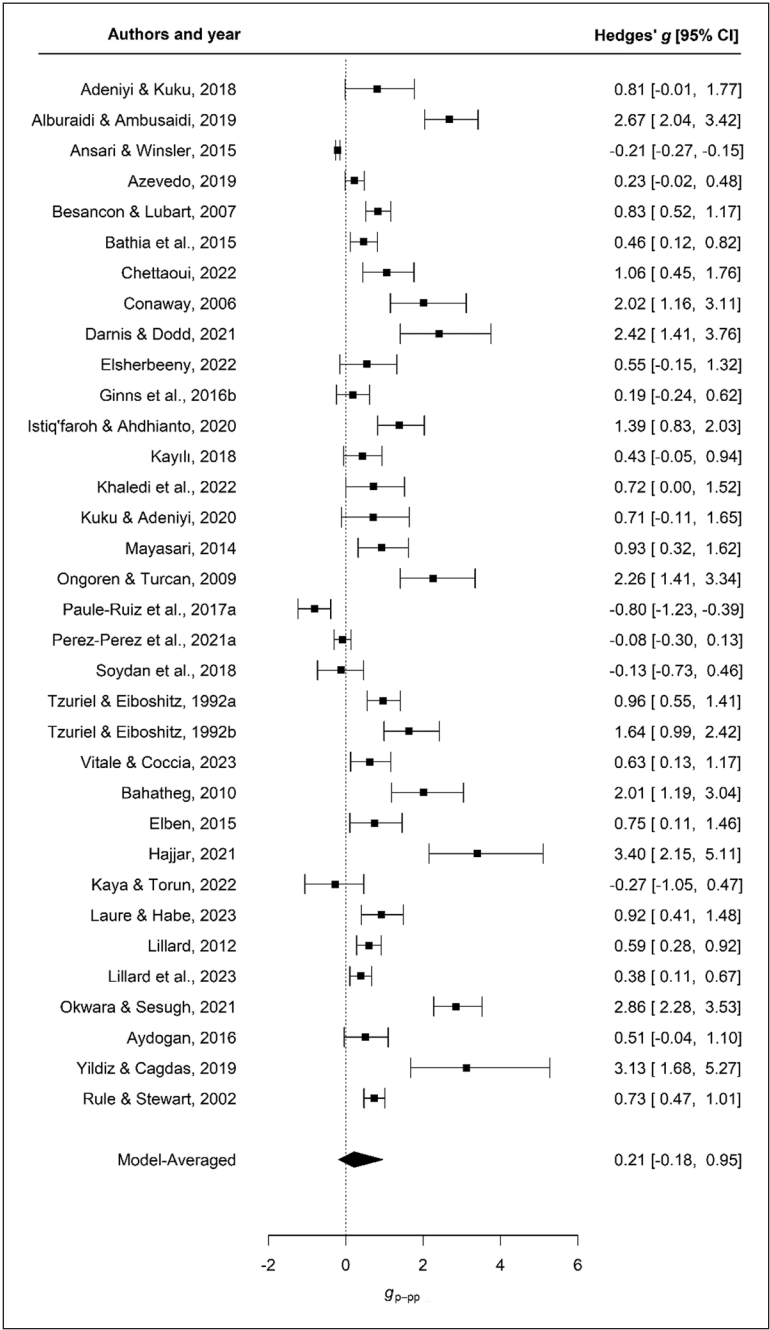


Figure 8. Forest plot for the robust Bayesian meta-analysis of only-post effect for the pretest-posttest ($\bar{g}_{p-pp}$).

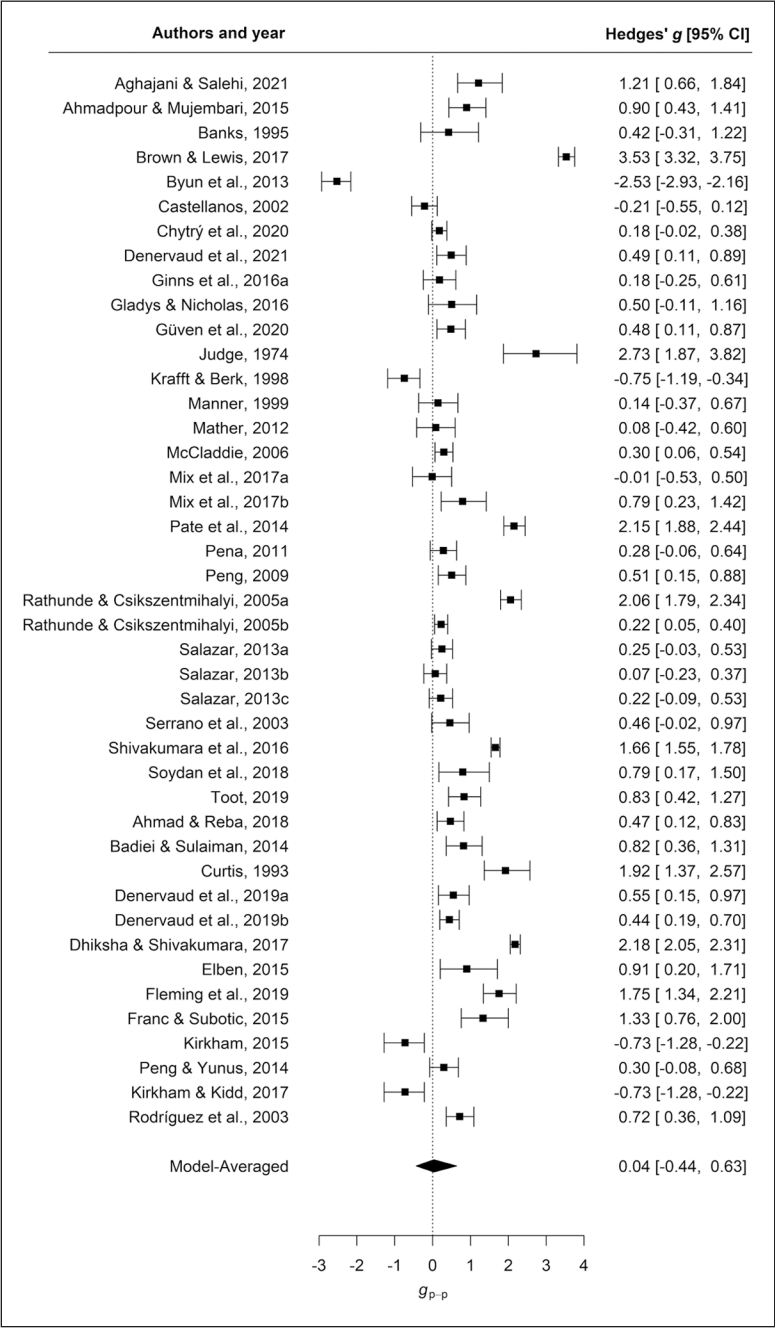


Figure 9. Forest plot for the robust Bayesian meta-analysis of only-post effect for the posttest-only design ($\bar{g}_{p-p}$).

In recent years, preregistration and the broader open-science movement have become key metascientific strategies for safeguarding against biased research practices. In meta-analysis, preregistration involves specifying and justifying objectives, inclusion criteria, and analytic decisions in advance, thereby ensuring transparency and reproducibility throughout the research process (Lakens et al., 2016; Moreau et al., 2022; Nosek & Lakens, 2014; Quintana, 2015). By clearly outlining goals and methods prior to data collection and analysis, preregistration minimizes researcher degrees of freedom and reduces the risk of bias or deviation during the review (Nosek et al., 2018).

As in other areas of empirical inquiry, meta-analyses can be used to formulate hypotheses grounded in previous literature, thus avoiding *HARKing*, the post-hoc construction of hypotheses after results are known (Kerr, 1998). In the present review, no a priori hypotheses were formulated. At the time the project began, no preregistered systematic reviews or preprints on the Montessori method were available, and none of the authors had conducted empirical research in this domain. Consequently, the review was designed as an exploratory synthesis, free from prior assumptions about direction or magnitude of effects. This open-ended approach enabled an unbiased assessment of Montessori education's impact on learning, structured around the following research questions.

RQ1: What Are the Main Characteristics of the Studies That Analyze the Use of the Montessori Method in Educational Contexts?

Our review revealed a substantial increase in research activity on the Montessori method in recent years (Figure 2). This growth likely reflects not only sustained societal interest in Montessori education but also more recent structural developments within the field, including the emergence of dedicated research networks, specialized journals, and targeted funding initiatives (e.g., A. Murray et al., 2023). At the same time, Montessori research remains only partially integrated into mainstream education research, with limited and uneven support for large-scale studies.

The present review identified 198 studies, considerably more than the 33 and 32 included in the recent reviews by Demangeon et al. (2023) and Randolph et al. (2023). Given the similar objectives and inclusion criteria among these works, potential detection or filtering biases may have led to the omission of relevant studies in earlier syntheses.

The distribution of research across educational stages was markedly uneven: 83.1% of studies focused on early childhood or primary education. This concentration is consistent with the Montessori method's historical prevalence in early schooling but also underscores a persistent limitation noted by Marshall (2017), who discussed the method's limited continuity beyond the primary level. Such discontinuity may leave students disoriented or excluded when transitioning from Montessori to conventional settings.

Most studies (82.8%) involved participants from the general population. In contrast, Sheppard et al. (2016) reported that Montessori-based interventions with individuals with cognitive deficits yielded benefits such as improved feeding behavior and limited cognitive or affective gains. This highlights the need to expand Montessori research to populations with specific educational or developmental needs, a domain that remains largely unexplored.

A notable proportion of the included studies were doctoral dissertations. In many cases, no subsequent peer-reviewed publications based on these dissertations could be identified, suggesting that a substantial part of the empirical work on Montessori education has not transitioned into the indexed journal literature. Geographically, nearly half of the research was conducted in the United States, reflecting a narrow concentration of empirical work on Montessori education. This result reaffirms criticism of sociocultural asymmetry in the implementation of the Montessori method (Canzoneri-Golden & King, 2023; Debs & Brown, 2017; Debs et al., 2022; Moquino et al., 2023).

Sample sizes varied substantially across studies. Among studies reporting total sample size, the mean N was 287 participants, whereas the median was 60 (range = 2–13,745), indicating a highly skewed distribution driven by a small number of very large studies. Most investigations relied on relatively small classroom-level samples. Such small-sample designs are associated with greater statistical instability and increased variability in effect size estimates, which may contribute to both inflated mean effects and substantial between-study heterogeneity observed in the literature. To further contextualize these constraints, we estimated the minimal detectable effect (MDE) under conventional assumptions ($\alpha = .05$, power = .80, two-group comparison with balanced allocation). Across studies reporting total sample sizes, the median MDE was approximately 0.72 (IQR = 0.51–1.06), indicating that the typical study would have been sufficiently powered primarily to detect relatively large effects. Under such conditions, smaller but potentially meaningful educational differences might remain undetected, and statistically significant findings could disproportionately reflect larger or more variable estimates. This structural feature of the literature may therefore contribute to both elevated average effects and the substantial between-study heterogeneity observed in our synthesis.

In terms of research design, posttest-only comparisons between Montessori and control groups were most prevalent, whereas fewer studies employed pretest–posttest control designs, which offer stronger causal inference. A recurring concern is that many investigations relied on comparisons between preexisting Montessori schools and conventional programs, without the research team having direct involvement in the implementation, monitoring, or documentation of the Montessori intervention. In such cases, researchers typically did not oversee how Montessori principles were enacted in practice, nor did they ensure comparability of instructional conditions across sites. This limited level of researcher control reduces replicability and introduces uncertainty regarding which variables drive the observed effects. As emphasized by Ferrero et al. (2021a, 2010b) and Román-Caballero et al. (2022), pretest–posttest designs with control groups, combined with clearer documentation of implementation conditions, are essential for robust evaluation of educational interventions.

In more than half of the studies, the researcher was also responsible for delivering the intervention. Although this information was often unreported, such overlap between instructor and investigator roles may introduce expectancy or allegiance biases when outcomes are assessed without independent procedures. At the same time, it is important to recognize the strong tradition of teacher-led and action research in both Montessori and conventional educational settings, where practitioners

systematically examine their own practice to improve instruction. The concern raised here is therefore not directed at practitioner inquiry per se, but at the implications of role overlap for causal inference and evidential robustness in studies aiming to establish comparative effectiveness. Greater transparency regarding role separation and outcome assessment procedures would strengthen interpretability in future research.

Finally, the overall empirical landscape of Montessori studies helps to contextualize several patterns observed in our synthesis. The scarcity of public Montessori programs with true random lotteries, the predominance of quasi-experimental designs with small samples, and the high variability in fidelity and outcome measurement collectively constrain internal validity and inflate between-study heterogeneity (Debs & Brown, 2017). These structural features provide the rationale for our emphasis on design-sensitive estimation and multimethod bias assessment in the subsequent analyses.

RQ2: What Is the Average Overall Effect on Learning of the Application of the Montessori Method in Educational Contexts?

Of the 198 studies included in this review, 114 provided sufficient data to compute effect sizes comparing the Montessori method with alternative instructional approaches. These studies employed a variety of experimental and quasi-experimental designs, including pretest–posttest and posttest-only formats with or without control groups. However, designs lacking control conditions cannot meaningfully compare learning outcomes between Montessori and non-Montessori contexts. Despite longstanding methodological guidance cautioning against such uncontrolled research (Campbell & Stanley, 1963; Knapp, 2016; Spurlock, 2018), they remain common in educational studies. Accordingly, our quantitative synthesis focused on those employing control-group designs (i.e., pretest–posttest and posttest-only) given their greater capacity to isolate intervention effects.

For pretest–posttest designs with a control group, the overall mean effect was significant ($\bar{g}_A = 1.03$, 95% CI [0.66, 1.40]), decreasing to $\bar{g}_A = 0.75$ [0.40, 1.08] after excluding outliers. For posttest-only designs, the pooled effect was likewise significant ($\bar{g}_P = 0.84$ [0.53, 1.14]) and remained so after excluding outliers ($\bar{g}_P = 0.80$ [0.56, 1.04]). A substantial number of studies yielded extreme or unstable values due to very small samples or analytic artifacts (see Appendix I). Reestimating pooled effects after removing these outliers was essential, as otherwise implausibly large effects could distort the true magnitude of the Montessori impact (Ferrero et al., 2021a; Viechtbauer & Cheung, 2010).

The influence of outliers can be illustrated by previous meta-analyses. Demangeon et al. (2023) examined five outcome domains (i.e., academic achievement, cognitive abilities, creative skills, motor skills, and social skills), reporting significant effects only for achievement ($g = 1.10$) and social skills ($g = 0.22$). However, closer inspection revealed that among 19 effects for academic achievement, 18 were nonsignificant and one was extreme ($g = 7.18$, 95% CI [3.75, 10.61]), disproportionately driving the overall result. Similarly, only two of 14 effects were significant for social skills. Randolph et al. (2023) found a similar pattern: reported effects were small ($g = 0.26$ [0.06, 0.46])

for academic ability; $g=0.17$ [0.03, 0.31] for language/literacy; $g=0.22$ [0.06, 0.39] for mathematics) and nonsignificant for science and social studies, yet again influenced by a few large effects. Such patterns illustrate how selective extreme values can inflate summary estimates.

Overall, our results suggest a nominally positive effect of Montessori education on student learning relative to comparison conditions, partly consistent with earlier reviews. However, our integrated quantitative and qualitative analyses challenge a simple interpretation of these effects. One recurrent concern is selection bias: Montessori students often come from higher socioeconomic backgrounds (Marshall, 2017), and socioeconomic status is strongly and consistently associated with academic achievement across educational systems (e.g., Reardon, 2011). In our review, baseline equivalence could be assessed in only 27 studies. While Montessori groups showed slightly higher pretest performance ($\bar{g}_{pre}=0.11$ [−0.01, 0.23], $p=.077$), these differences were not statistically significant, and similar trends appeared mainly in academic-achievement measures ($\bar{g}_{pre}=0.12$ [−0.01, 0.26], $p=.074$).

Another major source of distortion is publication bias, whereby studies reporting null or contradictory findings are less likely to appear in print. To address this, we adopted a multimethod approach (Román-Caballero et al., 2022), combining conventional diagnostics (funnel plots, Egger's test, and trim-and-fill procedures) with a robust Bayesian meta-analysis framework (e.g., Ciria et al., 2023). Unlike frequentist corrections, which rely on statistical significance and cannot quantify evidence for the absence of an effect, the Bayesian framework explicitly compares models with and without an effect, heterogeneity, and selection, weighting them by posterior probability. This approach provides a transparent account of model uncertainty and prevents over-interpretation of nonsignificant findings under high heterogeneity. The combined synthesis of this multimodal analysis showed strong evidence for bias and heterogeneity but only modest evidence for a genuine effect.

Based on this comprehensive evidence, we conclude that while the aggregated results superficially suggest superior performance among Montessori students, these apparent advantages are largely explained by publication bias, methodological weaknesses, and residual heterogeneity. Consequently, the mean effects reported in the literature should be interpreted cautiously: large observed differences most likely reflect the structural fragility of the evidence base rather than robust educational impacts. Once small-study and selection processes are modeled, support for a nonzero average effect of Montessori education becomes weak at best.

RQ3: What Is the Methodological Quality of the Studies That Analyze the Use of the Montessori Method in Educational Contexts, and Could These Characteristics Influence the Interpretations of the Effects Found?

Beyond the overall magnitude of effects, the credibility of Montessori research depends on methodological quality and transparency. Recent studies emphasize that assessing study quality is essential for interpreting systematic reviews and meta-analyses in education (Ferrero et al., 2021a, 2021b; García-Martínez et al., 2021). In our

review, we evaluated the methodological rigor of included studies using the scale developed by Ferrero et al. (2021a) and an additional instrument assessing fidelity of Montessori principles (Montessori, 2004). The Cochrane risk-of-bias scale is widely used in clinical research but not fully suited for educational contexts; hence, customized tools were more appropriate.

Our analysis revealed that only 22.78% of methodological quality criteria were met, while 40.11% were not met and 37.11% lacked sufficient information. This pattern reflects low overall quality and raises concerns about transparency and replicability. Experimental control was similarly limited: only 23.53% of studies randomized participants and 19.61% randomized groups. Information on blinding was scarce (93.63%–75.46% missing), posing a major potential bias (Boot et al., 2013). Merely 18.63% used active control groups, and 36.26% adjusted for pretest differences.

Educational interventions comparing experimental conditions with treatment-as-usual controls sometimes yield larger effects than those using active controls (Sala & Gobet, 2017). However, such differences are not necessarily artifacts of bias; they may arise from John Henry effects (where control participants modify performance when aware of the comparison) or overlapping pedagogical components in active controls (Boot et al., 2013). In our dataset, effect sizes were not larger for treatment-as-usual controls, suggesting that this concern, which was also observed in other meta-analyses, does not apply to the Montessori literature examined here. The few studies with pretest data did indicate a possible preexisting advantage for Montessori students, further complicating interpretation.

Transparency and open-science practices were extremely limited. Only two studies (0.98%) were preregistered and five (2.45%) openly shared their data. Such practices are essential for mitigating research biases (Kaplan & Irvin, 2015; Warren, 2018; Simmons et al., 2011) but remain virtually absent in this field. Without preregistration or open data, verification and cumulative synthesis become difficult.

Reporting deficiencies also affected generalization and replicability. Nearly 78.43% of studies did not provide replication details for the intervention, and 81.37% lacked information on the dependent variable. Only 34.80% of interventions met the standards proposed by Montessori (2004). Reliability and validity evidence were rarely reported (30.88% and 15.69%, respectively). The limited reporting of reliability and validity evidence may have important implications for the magnitude and variability of observed effects. Effects derived from non-standardized or researcher-developed measures, particularly in small samples, are typically more unstable and often larger than those obtained from normed or independently validated assessments. The scarcity of psychometric documentation in the Montessori literature, therefore, raises the possibility that part of the unusually large effect sizes and substantial heterogeneity observed in our synthesis reflect measurement imprecision rather than robust intervention impacts.

Regarding fidelity to Montessori principles, only 29.9%–35.78% of studies supplied sufficient information on adherence, whereas 60.29%–64.71% provided none. Although 73.53% employed Montessori curricular areas, reporting on principal implementation was inconsistent. Similarly to our intentions, attempts are currently being made to operationalize and validate instruments designed to assess fidelity associated

with the Montessori method, including the Montessori fidelity rubric (Culclasure et al., 2018), classroom observation tools (de Brouwer et al., 2024), and more recent efforts to develop and validate fidelity measures in early childhood and primary education contexts (A. K. Murray et al., 2025; Scippo, 2023), as well as empirical evidence linking fidelity to outcomes (Scippo, 2025).

Comparisons with prior reviews support these findings. Demangeon et al. (2023) also reported low quality scores, although numerical scales can oversimplify complex judgments by treating all criteria as equally weighted (Jüni et al., 1999). Randolph et al. (2023) described a generally low risk of bias but relied on instruments less adapted to education research, making their assessments less conclusive.

In sum, the methodological quality of Montessori studies is generally low, with limited adherence to both experimental standards and Montessori's own implementation criteria. This pattern, consistent with prior syntheses (Ferrero et al., 2021a, 2021b; García-Martínez et al., 2021), underscores the need for a critical reappraisal of methodological practices in educational intervention research.

It's important to note that our quality-stratified sensitivity analyses showed that effect magnitudes were not solely driven by methodological shortcomings. Studies rated as moderate or low in quality yielded comparable mean effects, suggesting that variability reflects structural features of the field, such as small samples (median $N=60$), quasi-experimental designs, limited psychometric documentation, and inconsistent fidelity reporting, rather than systematic inflation attributable to any single methodological flaw. Nevertheless, the lack of fully preregistered, blinded, and randomized trials continues to constrain evidential credibility and highlights the need for more rigorous methodological standards in future Montessori studies.

Finally, beyond methodological and statistical limitations, Montessori research is shaped by sociocultural asymmetries. Most available studies derive from Western, middle-class contexts, with minimal representation of marginalized or Global South populations. This imbalance restricts generalizability and reinforces structural inequities in the evidence base (Debs & Brown, 2017; Fallace, 2023; Moquino et al., 2023). Strengthening methodological rigor must therefore go hand in hand with diversifying research contexts and integrating equity and cultural responsiveness into empirical designs.

RQ4: Are There Factors That Could Moderate the Possible Heterogeneity of the Average Effect?

To explore sources of heterogeneity, we examined whether descriptive characteristics, methodological quality, and adherence to Montessori curricular principles moderated the observed effects (see Appendix L for details). Given the evidence of publication bias identified earlier, all moderator results should be interpreted with caution.

Among the examined moderators, only sample type significantly influenced effect heterogeneity. Significant effects were observed in typically developing populations, whereas effects for participants with disabilities were large but nonsignificant, and interventions including both populations yielded minimal effects. This pattern aligns partially with previous evidence suggesting potential benefits of Montessori approaches

for specific cognitive outcomes (Sheppard et al., 2016), though our data did not confirm statistically reliable effects for disabled populations.

We also analyzed outcome domain (academic achievement, cognitive abilities, social skills, emotional measures, creativity, and motor skills) as a categorical moderator (Appendix L, Table L1). Differences among these outcome types were not statistically significant ($F=0.32, p=.889$). Within the academic achievement category, which combined diverse indicators such as standardized tests, teacher-assigned grades, and researcher-developed tasks, considerable residual variability likely remains. The recent hypothesis proposing that variability in performance scores could better reflect individual growth among students in Montessori schools (Scippo, 2024) did not receive empirical support in our exploratory analyses, either overall or when accounting for the breadth of explicitly reported Montessori principles; however, the extent to which such variability may emerge under conditions of rigorously assessed high-fidelity implementation remains an open empirical question (see Appendix N).

Factors such as intervention agent and type of work approached significance. When interventions were implemented by researchers or jointly with teachers, effects were significant, whereas teacher-only interventions produced nonsignificant results. Notable differences emerged between research-led interventions and those co-led by researchers and teachers. In contrast, methodological quality did not significantly moderate effect heterogeneity. None of the Montessori principles showed significant moderation, although prepared environment and curricular areas approached significance, with larger effects found when these criteria were not reported as met.

Comparisons with previous syntheses are limited by sample-size differences. Demangeon et al. (2023) found no moderation for academic achievement but detected effects for social skills moderated by continent and study quality. That is, the effects were larger in Asian studies and smaller in studies of higher quality. Randolph et al. (2023) reported inconsistent findings due to insufficient effect sizes and weak moderator contrasts. The authors noted that they “were unable to reliably conduct a moderator and sensitivity analysis for each individual outcome because there was an insufficient number of effect sizes to do so” (p. 32), yet subsequently presented tables with subgroup results, leading to ambiguous interpretation.

Overall, our findings reveal few consistent moderators of Montessori outcomes. The limited and often nonsignificant moderation effects suggest that between-study variability likely arises from broader structural factors, such as design heterogeneity, small samples, and inconsistent reporting, rather than from any specific methodological or contextual moderator.

Limitations and Future Directions

Several features of the evidence base limit the strength and generalizability of our conclusions.

First, the pool of eligible studies remains dominated by small, quasi-experimental designs. Fully randomized trials are rare, and public Montessori lotteries are uncommon, restricting causal identification and long-term follow-up.

Second, fidelity to Montessori principles is variably implemented and inconsistently reported, reducing construct comparability across settings and potentially inflating or attenuating effects depending on local adaptations. Moreover, the very definition of what qualifies as “Montessori” remains heterogeneous, ranging from schools accredited by regional, national, or international Montessori organizations (e.g., AMI or AMS) to self-identified institutions that selectively adopt certain practices, making the treatment itself variable and only partially aligned with canonical Montessori features.

Third, outcome measures are highly heterogeneous and often lack reported psychometric properties, especially in nonacademic domains central to Montessori’s developmental aims.

Fourth, transparency practices such as preregistration and open sharing of data and materials remain uncommon, increasing uncertainty about analytic flexibility and selective reporting.

An additional procedural limitation concerns the screening process. Although full-text screening was conducted independently by two reviewers, title and abstract screening was performed by a single reviewer, which may increase the risk of selection bias despite subsequent full-text verification.

Finally, despite our multimethod approach to publication bias and design-sensitive modeling, residual small-study effects and unexplained heterogeneity persist.

Together, these limitations call for caution in interpretation and delineate a clear research agenda: larger samples, stronger assignment mechanisms, systematic fidelity documentation, standardized outcome batteries with reported reliability and validity, culturally responsive, and preregistered, openly shared workflows.

Conclusion

In conclusion, this preregistered systematic review and meta-analysis indicates that while the Montessori method shows promise for enhancing learning outcomes, the empirical evidence remains fragile. The positive average effects observed across studies are tempered by substantial concerns regarding methodological quality, publication bias, and inconsistent reporting. These findings emphasize the need for cautious interpretation and highlight the importance of strengthening transparency, replicability, and fidelity to Montessori’s pedagogical principles in future research.

Author Contributions

M.G.-G., y S.P.L.—literature search, data coding; A.A.L., M.G.-G., and S.P.L.—conceptualization; S.P.L.—data analysis; A.A.L., M.G.-G., and S.P.L.—writing—review & editing; S.P.L.—supervision.

ORCID iDs

Samuel P. León  <https://orcid.org/0000-0002-6980-2680>

Anastasiya A. Lipnevich  <https://orcid.org/0000-0003-0190-8689>

María García-Garrido  <https://orcid.org/0000-0002-6900-5470>

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by Junta de Andalucía under Grant HUM-642. MGG's participation was financed through a departmental collaboration grant from the University of Jaén awarded by the Ministerio de Educación y Formación Profesional (ref. 22CO1/001123).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

All the materials used throughout the review and analysis process are available through open and public access in the project created for this purpose in the OSF repository: <https://osf.io/j9v87/>.

Supplemental Material

Supplemental material for this article is available online.

References

- 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://psycnet.apa.org/doi/10.1037/amp0000389>
- Bartoš, F., & Maier, M. (2020). *RoBMA: An R package for Robust Bayesian meta-analyses* (R package version 3.1.0). <https://CRAN.R-project.org/package=RoBMA>
- Bartoš, F., Maier, M., Wagenmakers, E. J., Doucouliagos, H., & Stanley, T. D. (2023). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, 14(1), 99–116. <https://doi.org/10.1002/jrsm.1594>
- Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, 41(2), 257–278. <https://doi.org/10.1111/j.2044-8317.1988.tb00901.x>
- Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem of with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science*, 8, 445–454. <https://doi.org/10.1177/1745691613491271>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. <https://doi.org/10.1002/9780470743386>
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171–246). Rand McNally.
- Canzoneri-Golden, L., & King, J. (2023). Montessori education and critical race theory in the United States. In A. K. Murray, E. T. Ahlquist, M. K. McKenna, & N. Debs (Eds.), *The Bloomsbury handbook of Montessori education* (pp. 503–512). Bloomsbury.
- Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, 2(2), 115–144. <https://doi.org/10.1177/2515245919847196>

- Chen, A. (2021). Exploration of implementation practices of Montessori education in mainland China. *Humanities and Social Sciences Communications*, 8(1), 250. <https://doi.org/10.1057/s41599-021-00934-3>
- Ciria, L. F., Román-Caballero, R., Vadillo, M. A., Holgado, D., Luque-Casado, A., Perakakis, P., & Sanabria, D. (2023). An umbrella review of randomized control trials on the effects of physical exercise on cognition. *Nature Human Behaviour*, 7(6), 928–941. <https://doi.org/10.1038/s41562-023-01554-4>
- Coburn, K. M., & Vevea, J. L. (2019). *weightr: Estimating weight-function models for publication bias* (R package version 2.0.2). <https://cran.r-project.org/web/packages/weightr/>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
- Culclasure, B. T., Fleming, D. J., & Riga, G. (2018). *An evaluation of Montessori education in South Carolina's publics*. The Riley Institute at Furman University. <https://files.eric.ed.gov/fulltext/ED622145.pdf>
- de Brouwer, J., Morssink-Santing, V., & van der Zee, S. (2024). Validation of the teacher questionnaire of Montessori practice for early childhood in the Dutch context. *Journal of Montessori Research*, 10(1). <https://doi.org/10.17161/jomr.v10i1.21543>
- Debs, M. C., & Brown, K. E. (2017). Students of color and public Montessori schools: A review of the literature. *Journal of Montessori Research*, 3(1), 1–15. <https://doi.org/10.17161/jomr.v3i1.5859>
- Debs, M. C., de Brouwer, J., Murray, A. K., Lawrence, L., Tyne, M., & von der Wehl, C. (2022). Global diffusion of Montessori schools: A report from the 2022 global Montessori census. *Journal of Montessori Research*, 8(2), 1–15. <https://eric.ed.gov/?id=EJ1372150>
- Del Re, A. C., & Hoyt, W. T. (2014). *MAd: Meta-analysis with mean differences* (R package version 0.8-1). <http://cran.r-project.org/web/packages/MAd>
- Demangeon, A., Claudel-Valentin, S., Aubry, A., & Tazouti, Y. (2023). A meta-analysis of the effects of Montessori education on five fields of development and learning in preschool and school-age children. *Contemporary Educational Psychology*, 73, 102182. <https://doi.org/10.1016/j.cedpsych.2023.102182>
- Fallace, T. (2023). The racism of Maria Montessori. *Journal of Curriculum Studies*, 55(5), 619–631. <https://doi.org/10.1080/00220272.2023.2249067>
- Ferrero, M., Vadillo, M. A., & León, S. P. (2021a). A valid evaluation of the theory of multiple intelligences is not yet possible: Problems of methodological quality for intervention studies. *Intelligence*, 88, 101566. <https://doi.org/10.1016/j.intell.2021.101566>
- Ferrero, M., Vadillo, M. A., & León, S. P. (2021b). Is project-based learning effective among kindergarten and elementary students? A systematic review. *PLoS One*, 16(4), e0249627. <https://doi.org/10.1371/journal.pone.0249627>
- Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *robumeta: Robust variance meta-regression* (R package version 2.0). <https://cran.r-project.org/web/packages/robumeta/>
- García-Martínez, I., Gutiérrez Cáceres, R., Luque de La Rosa, A., & León, S. P. (2021). Analysing educational interventions with gifted students. Systematic review. *Children*, 8(5), 365. <https://doi.org/10.3390/children8050365>
- Harzing, A. W. (2007) *Publish or Perish*. <https://harzing.com/resources/publish-or-perish>
- 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>
- Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*, 282, 1054–1060. <https://doi.org/10.1001/jama.282.11.1054>

- Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects on large NHLBI clinical trials has increased over time. *PLoS One*, *10*, e0132382. <https://doi.org/10.1371/journal.pone.0132382>
- Kerr, N. L. (1998). HARKING: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Knapp, T. R. (2016). Why is the one-group pretest–posttest design still used? *Clinical Nursing Research*, *25*(5), 467–472. <https://doi.org/10.1177/1054773816666280>
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*, 1–10. <https://doi.org/10.1186/s40359-016-0126-3>
- L'Ecuyer, C. (2023). Montessori: origen y razones de las críticas a una de las pedagogas más controvertidas de la historia. *Revista Española de Pedagogía*, *81*(285), 270–270. <https://doi.org/10.22550/REP81-2-2023-01>
- Lillard, A., & Else-Quest, N. (2006). Evaluating Montessori education. *Science*, *313*(5795), 1893–1894. <https://doi.org/10.1126/science.1132362>
- Lillard, A. S. (2017). *Montessori: The science behind the genius*. Oxford University Press.
- Lillard, A. S., Tong, X., & Bray, P. M. (2023). Seeking racial and ethnic parity in preschool outcomes: An exploratory study of public Montessori schools vs. Business-as-usual schools. *Grantee Submission*, *9*(1), 16–36. <https://doi.org/10.17161/jomr.v9i1.19540>
- Lopata, C., Wallace, N. V., & Finn, K. V. (2005). Comparison of academic achievement between Montessori and traditional education programs. *Journal of Research in Childhood Education*, *20*(1), 5–13. <https://doi.org/10.1080/02568540509594546>
- Marshall, C. (2017). Montessori education: A review of the evidence base. *NPJ Science of Learning*, *2*(1), 1–9. <https://doi.org/10.1038/s41539-017-0012-7>
- Manship, K. (2023) Evaluating the efficacy of Montessori education. In A. K. Murray, E. T. Ahlquist, M. K. McKenna, & M. Debs (Eds.), *The Bloomsbury handbook of Montessori education* (pp. 209–218). Bloomsbury.
- Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *69*(5), 1091–1119. <https://doi.org/10.1111/rssc.12440>
- Montessori, M. (2004). *The Montessori method: The origins of an educational innovation: Including an abridged and annotated edition of Maria Montessori's the Montessori method*. Rowman & Littlefield Publishers.
- Moquino, T., Walker, N., & Kitchens, K. (2023). Beyond authenticity: Indigenizing Montessori education in settler colonial United States. In A. K. Murray, E. T. Ahlquist, M. K. McKenna, & M. Debs (Eds.), *The Bloomsbury handbook of Montessori education* (pp. 513–524). Bloomsbury Academic.
- Moreau, D., & Gamble, B. (2022). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods*, *27*(3), 426–432. <https://doi.org/10.1037/met0000351>
- Morris, S. B. (2008). Estimating effect sizes from pretest–posttest–control group designs. *Organizational Research Methods*, *11*, 364–386. <https://doi.org/10.1177/1094428106291059>
- Murray, A., Debs, M., McKenna, M., & Ahlquist, E. M. T. (2023). *The Bloomsbury handbook of Montessori education*. Bloomsbury Academic.
- Murray, A. K., & Daoust, C. (2023). Fidelity issues in Montessori research. In A. K. Murray, E. T. Ahlquist, M. K. McKenna, & M. Debs (Eds.), *The Bloomsbury handbook of Montessori education* (Vol. 20, pp. 199–208). Bloomsbury Academic.

- Murray, A. K., Murray, A. J., Daoust, C. J., & Gerker, H. E. (2025). Developing a tool to evaluate early childhood education implementation fidelity: Measure of Montessori Implementation pilot study results. *Frontiers in Education*, 10, 1603908. <https://doi.org/10.3389/educ.2025.1603908>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606. <https://doi.org/10.1073/pnas.1708274114>
- Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, 45(3), 137–141. <https://doi.org/10.1027/1864-9335/a000192>
- 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. *BMJ*, 372. <https://doi.org/10.1136/bmj.n71>
- Pustejovsky, J. (2024). *clubSandwich: Cluster-Robust (Sandwich) variance estimators with small-sample corrections* (R package version 0.5.10.9999). <http://jepusto.github.io/club-Sandwich/>
- Pustejovsky, J. E., & Rodgers, M. A. (2019). Testing for funnel plot asymmetry of standardized mean differences. *Research Synthesis Methods*, 10(1), 57–71. <https://doi.org/10.1002/jrsm.1332>
- Quintana, D. S. (2015). From pre-registration to publication: A non-technical primer for conducting. *Frontiers in Psychology*, 6, 167512. <https://doi.org/10.3389/fpsyg.2015.01549>
- Randolph, J. J., Bryson, A., Menon, L., Henderson, D. K., Kureethara Manuel, A., Michaels, S., Rosenstein, D. L. W., McPherson, W., O’Grady, R., & Lillard, A. S. (2023). Montessori education’s impact on academic and nonacademic outcomes: A systematic review. *Campbell Systematic Reviews*, 19(3), e1330. <https://doi.org/10.1002/cl2.1330>
- Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G. J. Duncan & R. J. Murnane (Eds.) *Whither opportunity?: Rising inequality, schools, and children’s life chances* (pp. 91–116). Russell Sage Foundation. <https://doi.org/10.7758/9781610447515.10>
- Román-Caballero, R., Vadillo, M. A., Trainor, L. J., & Lupiáñez, J. (2022). Please don’t stop the music: A meta-analysis of the cognitive and academic benefits of instrumental musical training in childhood and adolescence. *Educational Research Review*, 35, 100436. <https://doi.org/10.1016/j.edurev.2022.100436>
- Rosenthal, R. (1991). Applied social research methods series (rev. ed.). In *Meta-analytic procedures for social research* (Vol. 6). Sage Publications, Inc.
- Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science*, 26, 515–520. <https://doi.org/10.1177/0963721417712760>
- Scippo, S. (2022). The recent growth and current dissemination of the Montessori primary school classes in Italy. *STUDIUM EDUCATIONIS-Rivista semestrale per le professioni educative*, 1, 004–017. <https://doi.org/10.7346/SE-012022-01>
- Scippo, S. (2023). Costruzione e validazione di uno strumento per misurare le pratiche educative Montessori nella scuola primaria italiana [Construction and validation of a tool to measure Montessori educational practices in the Italian primary school]. *Journal of Educational, Cultural and Psychological Studies (ECPS)*, 28, 117–135. <https://doi.org/10.7358/ecps-2023-028-scis>

- Scippo, S. (2024). Montessori primary schools' effectiveness: A quasi-experimental study on schooling outcomes. *School Effectiveness and School Improvement*, 35(2), 193–213. <https://doi.org/10.1080/09243453.2024.2349537>
- Scippo, S. (2025). Montessori education outcomes in Italy: Assessing the influential role of duration and implementation fidelity. *Issues in Educational Research*, 35(3), 1197–1220. <http://www.iier.org.au/iier35/scippo.pdf>
- Scippo, S., & Cañigual-Viñals, D. (2023). La difusió actual de l'aplicació escolar de la pedagogia Montessori a Catalunya. *REIRE Revista d'Innovació i Recerca en Educació*, 16(2), 1–13. <https://doi.org/10.1344/reire.40985>
- Sheppard, C. L., McArthur, C., & Hitzig, S. L. (2016). A systematic review of Montessori-based activities for persons with dementia. *Journal of the American Medical Directors Association*, 17(2), 117–122. <http://dx.doi.org/10.1016/j.jamda.2015.10.006>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Spurlock, D. R. (2018). The single-group, pre-and posttest design in nursing education research: It's time to move on. *Journal of Nursing Education*, 57(2), 69–71. <https://doi.org/10.3928/01484834-20180123-02>
- 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., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, 40(6), 604–634. <https://doi.org/10.3102/1076998615606099>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <http://www.jstatsoft.org/v36/i03/>
- Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1(2), 112–125. <https://doi.org/10.1002/jrsm.11>
- Warren, M. (2018). *First analysis of 'pre-registered' studies shows sharp rise in null findings*. <https://www.nature.com/articles/d41586-018-07118-1>

Author Biographies

Samuel P. León, PhD, is associate professor in the Department of Pedagogy, Research Methods and Diagnosis in Education at the University of Jaén, España; email: sparra@ujaen.es. His research interests are related to research methodology in educational contexts, applying the principles of psychometrics to education research, and above all, trying to improve education research through meta-science.

Anastasiya A. Lipnevich, PhD, is full professor of educational psychology at Queens College and the Graduate Center of the City University of New York and an inaugural visiting scholar at the National Board of Medical Examiners; email: a.lipnevich@gmail.com. Her research interests include: instructional feedback, formative assessment, alternative ways of cognitive and non-cognitive assessment, and the role of psychosocial characteristics in individuals' academic and life achievement.

María García-Garrido is a PhD candidate and research associate at the Department of Pedagogy, Research Methods and Diagnosis in Education at the University of Jaén, España; email: mggarrid@ujaen.es. His research interests are focused on conducting metascientific actions in educational interventions.