Cogent Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/oaed20

# A systematic review to evaluate the risk of bias of meta-analyses reporting experimental educational interventions focused on academic performance

**Micaela Sánchez-Martín, Marta Gutiérrez-Sánchez, Eva María Olmedo-Moreno & Fernando Navarro-Mateu**

**To cite this article:** Micaela Sánchez-Martín, Marta Gutiérrez-Sánchez, Eva María Olmedo-Moreno & Fernando Navarro-Mateu (2024) A systematic review to evaluate the risk of bias of meta-analyses reporting experimental educational interventions focused on academic performance, Cogent Education, 11:1, 2385785, DOI: 10.1080/2331186X.2024.2385785

**To link to this article:** https://doi.org/10.1080/2331186X.2024.2385785

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

View supplementary material

Published online: 05 Aug 2024.

Submit your article to this journal

Article views: 221

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=oaed20

COGENT EDUCATION
2024, VOL. 11, NO. 1, 2385785
https://doi.org/10.1080/2331186X.2024.2385785

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

OPEN ACCESS Check for updates

# A systematic review to evaluate the risk of bias of meta-analyses reporting experimental educational interventions focused on academic performance

Micaela Sánchez-Martín[a], Marta Gutiérrez-Sánchez[b], Eva María Olmedo-Moreno[c] and Fernando Navarro-Mateu[d,e]

[a]Departamento de Métodos de Investigación y Diagnóstico en Educación, Universidad de Murcia, Murcia, Spain; [b]Departamento de Métodos de Teoría e Historia de la Educación, Universidad de Murcia, Murcia, Spain; [c]Departamento de Métodos de Investigación y Diagnóstico en Educación, Universidad de Granada, Granada, Spain; [d]Departamento de Metodología y Psicología Básica, Universidad de Murcia, Murcia, Spain; [e]Unidad de Docencia, Investigación y Formación en Salud Mental (UDIF-SM), Servicio Murciano de Salud, Murcia, Spain

## ABSTRACT

**Introduction:** Concerns about the risk of bias (RoB) of Meta-analysis (MAs) have grown in parallel with the exponential increase in the number of publications in science. However, this has not been properly assessed in Education. The aims were to evaluate the RoB of MAs in Education and to identify potential predictors of a lower RoB.

**Methods:** Systematic review. Selection criteria were all MAs of experimental design evaluating the effectiveness of educational interventions on any academic outcome published from 1 January 2009 (year of publication of the first PRISMA guideline) published in English or Spanish, with the exclusion of those with other designs, evaluating other outcomes or not accessible to full text. A systematic search was performed in four databases (ERIC, Web of Science, Scopus, and PubMed) until March 2022. A preregistered protocol was used to extract data on study characteristics, PRISMA compliance, and RoB, based on the AMSTAR 2 instrument, and dichotomized as low *vs.* high RoB. The study selection and data extraction process were independently conducted by two researchers and disagreements were solved by consensus or by a third researcher. Statistical analysis: A flow-diagram and descriptive tables were tabulated. As a measure of association, odds ratios (OR) and its 95% confidence intervals were calculated by logistic regression analysis with dichotomized RoB as the dependent variable.

**Results:** A total of 69 meta-analyses of studies were identified. Almost 90% ($n = 62$) of them were rated with a critically low overall confidence level, and almost 70% ($n = 49$) had a high RoB. Factors related to a low RoB were adherence to PRISMA guideline (OR = 5.5; 95%CI: 1.8–16.6), the most recent studies (OR = 7.4; 95%CI: 1.5–35.3), a higher number of authors (OR = 1.4; 95%CI: 1.1–1.9), a corresponding author from a European country (OR = 3.7; 95%CI: 1.1–12.8), and publishing in the health educational area (OR = 13.4; 95%CI: 3.6–49.6).

**Conclusions:** Our study raises concerns regarding the methodological quality of published MA in Education. The use of instruments, such as AMSTAR 2 and PRISMA 2020, may improve the quality of future MA in Education.

## PUBLIC INTEREST STATEMENT

Meta-analyses and systematic reviews are research designs developed to identify, evaluate, and summarize the findings of all relevant individual studies over a given research question to make the best available evidence more accessible to readers and decision-makers. It is crucial to systematically assess their methodological weaknesses that could diminish the confidence in their results (risk of bias) and to identify related factors. We assessed the risk of biases in meta-analyses of experimental interventions developed in Education to improve academic performance. Our results suggested that the overall confidence in the results of most of the included studies was low or critically low and several predictors of lower risk of bias were identified (e.g. most recent

## ARTICLE HISTORY

Received 5 March 2024
Revised 29 May 2024
Accepted 11 July 2024

## KEYWORDS

Systematic review; meta-analysis; AMSTAR-2; PRISMA; educative interventions; academic outcome

## SUBJECTS

Classroom Practice; Multicultural Education; Research Methods in Education; Education-Social Sciences

CONTACT Eva María Olmedo-Moreno emolmedo@ugr.es Departamento de Métodos de Investigación y Diagnóstico en Educación, Universidad de Granada, Granada, Spain

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

publications, higher number of authors, European corresponding author and published in education in health sciences). Authors, editors, and users should be aware of the importance of improving the quality of the methodology of these designs.

## 1. Introduction

Systematic reviews (SRs) and meta-analyses (MAs) were developed to identify, evaluate, and synthesize the results of different research studies, selected through a systematic search, to answer a specific and focused research question and they have been suggested to be the highest form of published evidence available to professionals (Guyatt et al., 2000). The term MA was first introduced by Gene Glass in the field of education in 1976 (Glass, 1976; Glass et al., 1981) and it quickly spread in medicine, psychology, and, more recently, in other areas of knowledge (e.g. Education). Nowadays, the increasing number of SRs and MAs published has been exponential in all areas of knowledge (Bastian et al., 2010; Gurevitch et al., 2018). This impressive growth has been accompanied by a growing concern for the quality of their methodology (Anguera, 2023). In this context, different studies have reviewed the quality of published SRs and MAs, and their results suggest that, too often, published MAs tend to be of low quality in different knowledge areas, such as Medicine (Bidhendi Yarandi et al., 2021; Delaney et al., 2005; Hameed et al., 2020; Huedo-Medina et al., 2016) or Psychology (Cafri et al., 2010; López-Nicolás et al., 2022). In Education, to the best of our knowledge, only two studies have analyzed the risk of bias (RoB) of published MAs showing that MAs often present methodological deficiencies that may limit the validity of their conclusions (Ahn et al., 2012; Eser & Yurtçu, 2020). Both used different scales specifically developed by the authors themselves limiting the comparability of the results, and the most recent is limited to MAs published in educational sciences in Turkey from 2010 to 2019.

To systematize the assessment of the RoB in MAs, different instruments have been developed, such as the AMSTAR-2 (A MeaSurement Tool to Assess systematic Reviews) (Shea et al., 2017). This instrument is the up-dated version of the original AMSTAR, a measurement tool developed to assess the methodological quality of systematic reviews (Shea, Grimshaw, et al., 2007). Different studies have used this instrument to assess the methodological quality of SRs and MAs in the field of health education (Hume et al., 2023; Jin et al., 2023; Martinez-Calderon et al., 2023), and all of them call for a necessary improvement in the quality of this type of research due to the low level found. We are not aware of the publication of any review of MAs of primary experimental studies in the education field with the assessment of the RoB with this structured instrument that allow comparisons with other knowledge scientific areas.

This concern about the low quality of SRs and MAs contributed to the initially development of the QUOROM (QUality Of Reporting Of Meta-analysis) statement with the aim of helping to improve the quality and transparency of the report of MAs of experimental studies (Moher et al., 1999). Subsequently, it was revised to incorporate some improvements in the publication of SRs and MAs of experimental and observational studies and changed to the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analysis) statement (Moher et al., 2009). It has recently been updated to PRISMA 2020 statement (Page, McKenzie, et al., 2021). This instrument consists of a 27-item checklist reporting recommendations to help researchers properly report their SR. Although the main goal was to benefit authors, editors, peer reviewers, and different users of SRs, such as guideline developers, policy makers, and other stakeholders by improving the transparency and quality of reporting. However, PRISMA 2020 was not developed to assess the risk of bias assessment of SRs as other specific tools have already been developed, such as AMSTAR 2, among others. The use of PRISMA in Education has been increasingly growing in last years (Holmqvist & Ekström, 2024; Moreno et al., 2022). The endorsement of PRISMA statement when writing a manuscript has been associated with a lower RoB and higher methodological quality (Panic et al., 2013; Sharma & Oremus, 2018).

The assessment of the quality of a study and its RoB are often used as synonyms. However, they are not completely the same. In its latest version, the PRISMA 2020 guideline distinguishes quality from RoB assessment (Page, Moher, et al., 2021). The former is a broader concept and often includes constructs beyond those that may bias study results (e.g. ethical issues, applicability, or imprecise explanations).

While the latest, the RoB assessment, refers to those aspects of the methodology (specifically related to the design, development, or the analysis) that have the potential to systematically bias the results. It focuses specifically on the internal validity of the study or the degree of confidence that the study results have been obtained and interpreted appropriately.

Given the increasing importance of MAs in synthesis the evidence and their use in professional decision making, understanding factors affecting their quality are determinant in Education. There were three main aims of this study. First, to describe the RoB of MAs of experimental studies using the AMSTAR 2 checklist. A previous decision was made to focus on the specific topic of academic performance as the outcome of the interventions. Second, to assess whether the use of the PRISMA guideline is related to a lower RoB. And, finally, to identify potential predictors of a lower RoB in MAs. The main research question according to the PICO framework (P: Population; I: Intervention; C: Comparator; and O: Outcome) (Page, Moher, et al., 2021) would be: Do MAs of experimental studies in the field of education (P) that adhere to the PRISMA guidelines (I) have a lower RoB (O) compared to those that do not adhere to PRISMA guidelines (C)?

## 2. Methods

### 2.1. Protocol and registry

A methodological systematic review, a special type of umbrella review (López-López et al., 2022), was performed to assess the RoB of MAs of educational interventions focused on different academic outcomes was performed. The protocol was registered in OSF (Open Science Framework: https://doi.org/10.17605/OSF.IO/PSKN6). This report has been written according to the PRISMA 2020 statement (Page, McKenzie, et al., 2021).

### 2.2. Study eligibility criteria

Inclusion criteria were as follows: (i) MAs of experimental designs evaluating the effectiveness of educational interventions on any academic outcomes performance; (ii) studies wrote in English or Spanish; and (iii) published from 1 January 2009, corresponding to the publication date of the first PRISMA guideline (Moher et al., 2009) until 30 April 2022. Exclusion criteria were: (i) MAs based on non-experimental designs, such as single-group pre- and post-tests with no control group, qualitative studies, and other designs examining only the correlation between the intervention and the student achievement, as well as meta-analyses evaluating the effect of educational interventions on other outcomes apart from academic performance; (ii) traditional narrative reviews or other publication types such us editorials, book chapters, conferences/posters, protocols or theoretical descriptions of educational interventions; and (iii) articles not accessible to full text.

### 2.3. Information sources, search strategy, and study selection

Comprehensive electronic searches were performed using the following databases: ERIC (Education Resources Information Center), WOS (Web of Science), SCOPUS, and PubMed from 1 January 2009 until 30 April 2022. The following search terms were used: 'meta-analysis' AND 'educational intervention' AND 'academic outcome'. See the description of each of the specific database search strategies in Appendix A. The reference lists of the included MAs were also manually examined searched to identify other potentially eligible studies. No restrictions were placed on sample size or ethnicity.

Identified documents were entered with Zotero (https://www.zotero.org/), an open-access reference manager tool, and duplicates were manually deleted. The selection process was performed in several phases. First, an initial screening search was carried out to eliminate duplicates and documents clearly not related to the subject of study. Second, titles and abstracts were independently reviewed by two reviewers for inclusion and exclusion criteria. Finally, full-text documents were independently reviewed again by the same two independent reviewers. In case of disagreement, a consensus was reached by the reviewers, or a decision was made by a third researcher.

### 2.4. Data extraction

The following data were extracted from each included study following a previously defined protocol: first author, year of publication [dichotomized in 2009–2016 *vs.* 2017–2022, according to the year of publication of the AMSTAR 2 in 2017 (Shea et al., 2017)], journal, number of authors, adherence to the PRISMA guidelines and journal impact factor, corresponding author's countries and their aggrupation in continents, search strategy (regional or national/international), area of education (education, psychology or health), type of intervention (categorized as methodological strategies, organizational strategies, physical activities and other or a combination of the previous), characteristics of the target population according to educational level involved (Primary, secondary school, higher education, a combination of the formers all levels included or) and the RoB assessment. The data extraction process was carried out independently by two researchers using a previously elaborated protocol. Disagreements were resolved by consensus or by the intervention of a third evaluator.

### 2.5. Risk of bias (RoB) assessment

To assess the RoB of the selected MAs, the AMSTAR 2 (A MeaSurement Tool to Assess systematic Reviews) was applied. A complete description of the instrument, rationale, and further information is described elsewhere (Shea et al., 2017). Briefly, this tool was specifically designed to assess the RoB of SRs and MAs of experimental and non-experimental studies. AMSTAR 2 comprises 16 items for which 'yes (Y)' or 'no (N)' judgments can be applied. For five items (2, 4, 7, 8, 9) an additional response is added, 'partially yes (PY)'. Seven of the 16 items are considered to be critical: 'development of the study protocol' (item 2); 'comprehensiveness of the literature search strategy' (item 4); 'providing a list of excluded studies with reasons' (item 7); 'appropriate assessment of the RoB of individual included studies' (item 9); 'use of appropriate meta-analytical methods' (item 11); 'consideration of RoB when interpreting and discussing the results' (item 13); and 'assessment of the presence of publication bias and discussion of its impact on the results' (item 15). Subsequently, the remaining nine items are classified as non-critical (for a complete list of all items see the foot note in Figure 2). The original rating is not intended to generate an overall score, but allows an overall judgment of the confidence in the results of the target review in four categories according to the following suggested criteria: 'high' (no major critical flaws or only one flaw within non-critical items); 'moderate' (no major critical flaws and more than one flaw in non-critical items); 'low' (one critical flaw with or without non-critical flaws); or 'critically low' (more than one critical flaw with or without non-critical flaws). However, the original authors themselves suggest that users may prefer to define other variations in the proposed evaluation system or to incorporate some modifications according to future researcher's aims (Shea et al., 2017). Accordingly, a decision to adopt a less restrictive and more conservative evaluation criterion, based exclusively on the number of critical domains and independently of the number of non-critical items, was adopted and a dichotomized RoB assessment was defined as follows: 'low risk' (up to two critical flaws) and 'high risk' (three or more critical weaknesses). This decision was based on an expected very low rate of studies rated as high or moderate according to previous results (Ahn et al., 2012; Eser & Yurtçu, 2020).

### 2.6. Statistical analyses

Inter-rater agreement in both, the selection, and data extraction process, was measured by Cohen's Kappa index. A flow-chart, according to PRISMA 2020 guidelines (Page, McKenzie, et al., 2021) was used. A description of the individual characteristics of included studies and a list of excluded studies with their reason for exclusion was tabulated. Results of the RoB assessment will be presented in two complementary forms, tabulated by each included studies by the items of the AMSTAR 2 and as a figure with the summary results of each individual item (Higgins et al., 2022). An initial descriptive analysis of categorical (percentage) and continuous variables (means and standard deviations) was performed. Chi-squared test or Fisher's exact test was used to analyze differences according to the high *vs.* low RoB. As a measure of association, the odds ratio (OR) and its 95% confidence interval (95%CI) with its *p*-value were calculated by simple logistic regression analysis with RoB as the

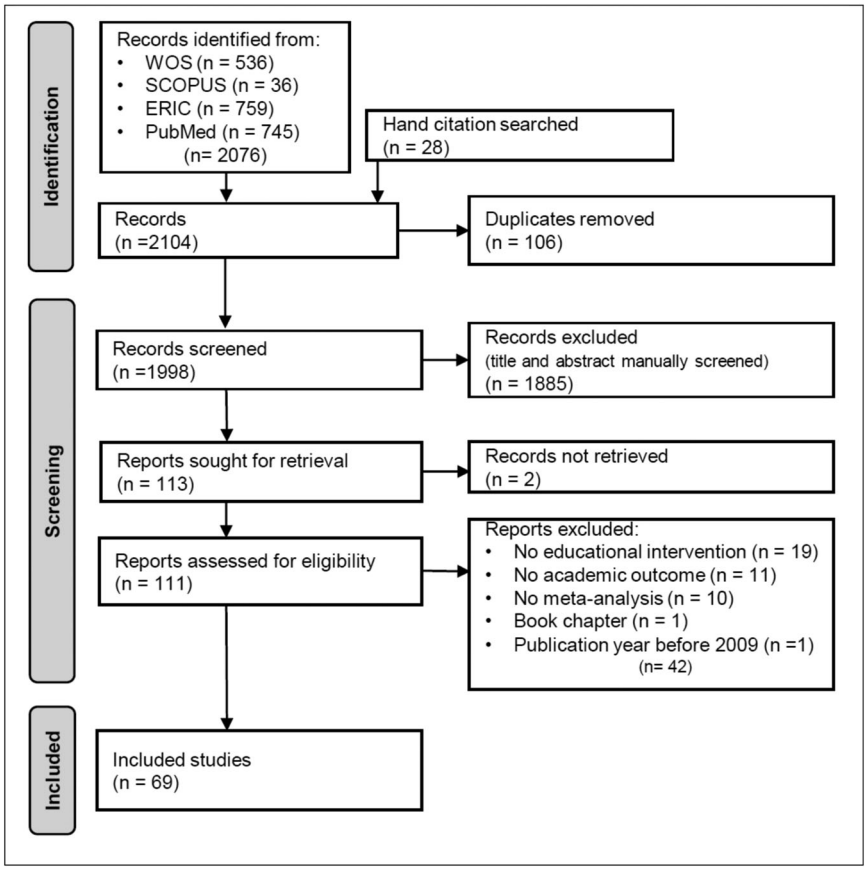


**Figure 1.** PRISMA 2020 flow diagram (adapted from Page, McKenzie, et al., 2021).

dependent variable. Bilateral statistical tests were used with an alpha level of .05. Statistical analyses were performed using SPSS software (V.28.0).

## 3. Results

### *3.1. Study eligibility and data collection*

A total of 2076 studies were initially identified. After the selection process (see flow-diagram in Figure 1), a final sample of 69 studies were included. The list of excluded articles after a full-text assessment with the reason for exclusion and the description of included studies are described in supporting information (Appendices B and C, respectively). The median (SD) of the Cohen's kappa inter-rater agreement coefficient in the selection process was 0.87 (0.20) and ranged from 0.59 to 1.00. Table 1 describes the main variables in detail. Only 24 studies (34.8%) describe an explicit use of the PRISMA guideline. Most studies ($n = 47$, 67.1%) were published after 2017, when the AMSTAR 2 was published, in journals with a Q index of 1 (41, 73.2%) and a median impact factor of 2.2 ($SD = 1.8$). The mean number of authors was 3.7 ($SD = 2.6$) with a minimum of one author in seven studies (10.1%), 16 studies (22.9%) with five or more authors, and a maximum of 17 in one study (1.4%) (see Appendix C). Only a minority of studies used a geographical limitation in their search strategy (13, 18.8%). The first three countries from which the corresponding authors were from were United States (18, 26.1%), Turkey (12, 17.4%), and China (11, 15.9%) and, when grouped in continents, Australia and Asia were the most frequent (32, 46.4%). Most of the studies were published in journals of education (43, 62.3%), focused on methodological interventions (30, 43.5%), and with a combination (24, 34.8%) or with the inclusion of all levels (19, 27.5%) of target population (mainly primary, secondary and higher education). The median (*SD*) of the Cohen's kappa inter-rater agreement coefficient was 0.90 (0.19) for the general data extraction process and 0.75 (0.29) for the domains of the AMSTAR 2.

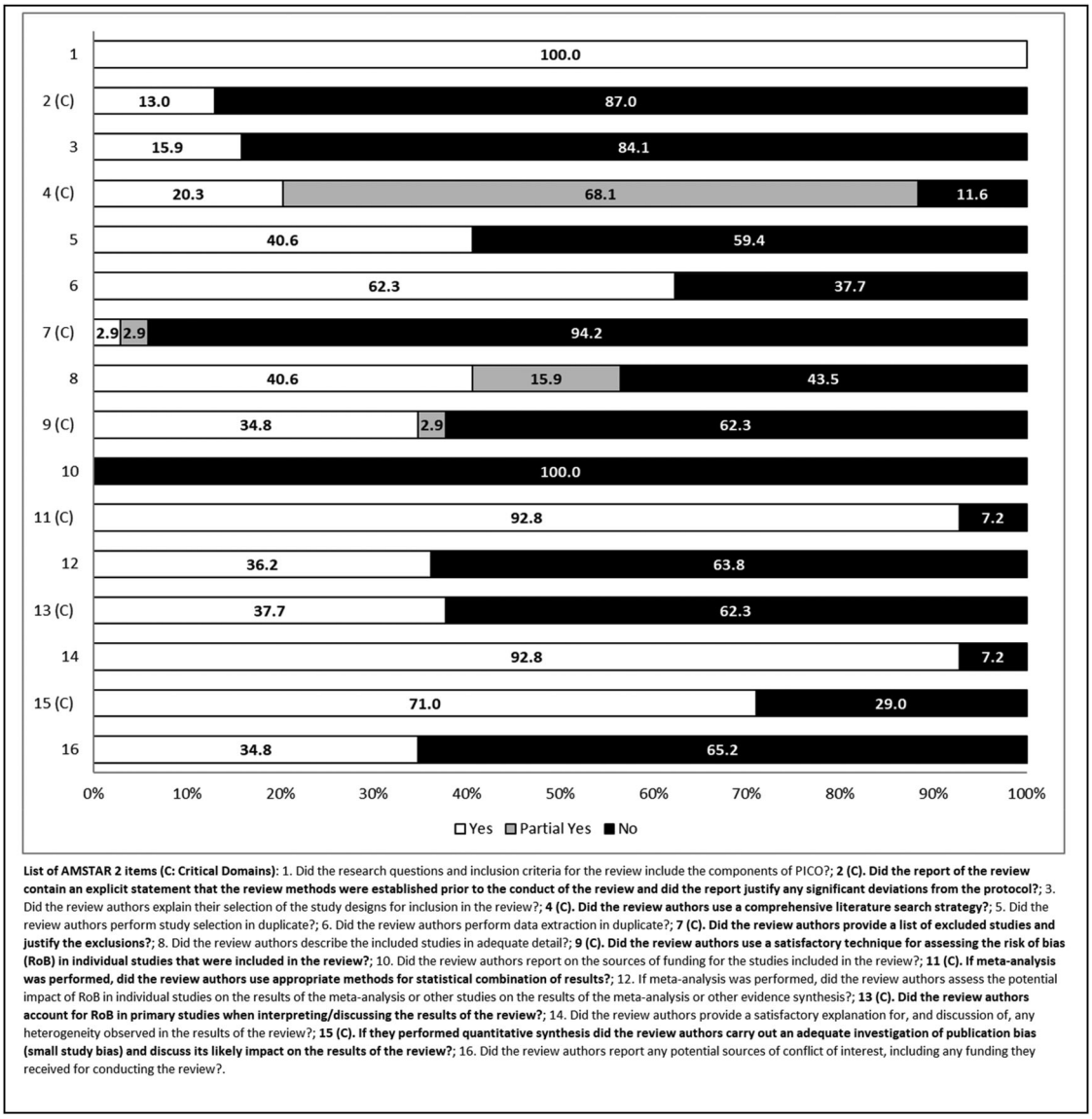


**Figure 2.** Summary (percentage) results of the risk of bias assessment of the included meta-analyses ($n = 69$) assessed by the AMSTAR 2 (A MeaSurement Tool to Assess Systematic Reviews) instrument.

### *3.2. Risk of bias (RoB) analyses*

The RoB of included studies was assessed with the AMSTAR 2. A table with the individual assessment of each included study is presented in supporting information (Appendix D) and Figure 2 represents the summary results of each individual item. Among the seven critical domains (see Figure 2), four of them were not completed in more than 30% (see Figure 2): 94.2% of the MAs did not include a list of excluded studies with the reason for exclusion (item 7), 87.0% did not explicitly include a statement that the review methods were established before the conduct of the field work (item 2), and 63.2% did not use a satisfactory technique for individual assessment of the RoB in the included studies (item 9) and did not account for the RoB when interpreting or discussing the results (item 13). The remaining three studies were not completed in <30%: to use a comprehensive literature search strategy (item 4), 11.6%; to use appropriate statistical methods to combine the results (item 11) 7.2%; and to evaluate the publication bias and to discuss its potential impact on the results (item 15) 29.0%. Among the non-critical items, the two more frequent flaws were not reporting sources of funding for the included studies (100.0%) and not explaining their selection of the study designs for inclusion (84.1%). Almost 90% ($n = 62$) of the studies were evaluated with a critically low overall confidence in the results due to more

**Table 1.** General description of variables.

| Variables | *n* | % | Variables | *n* | % |
|---|---|---|---|---|---|
| Total | 69 | | | | |
| PRISMA adherence | | | Continents | | |
| No | 45 | 65.2 | Australia and Asia | 32 | 46.4 |
| Yes | 24 | 34.8 | Europe | 18 | 26.1 |
| Number of authors (mean, *SD*) | 3.7 | 2.6 | North-America | 19 | 19.0 |
| Year of publication (mean, *SD*) | 2017.5 | 2.9 | Area of education | | |
| 2009–2016 | 23 | 32.9 | Education | 43 | 62.3 |
| 2017–2022 | 47 | 67.1 | Psychology | 8 | 8.0 |
| Impact factor (mean, *SD*) | 2.2 | 1.8 | Health | 18 | 18.0 |
| Search strategy with geographic restrictions | | | Type of intervention | | |
| Regional/national | 13 | 18.8 | Methodological strategies | 30 | 43.5 |
| International | 56 | 81.2 | Organizational strategies | 17 | 24.6 |
| Corresponding author's country | | | Physical activities and others | 10 | 14.5 |
| UUEE | 18 | 26.1 | Combination | 13 | 18.6 |
| Canada | 1 | 1.4 | Target population | | |
| UK | 6 | 8.7 | Primary, secondary | | |
| The Netherlands | 6 | 8.7 | school and disabilities | 10 | 10.0 |
| Portugal | 1 | 1.4 | Higher education | 15 | 21.7 |
| Norway | 1 | 1.4 | All levels included | 19 | 27.5 |
| Finland | 1 | 1.4 | Combination of levels | 24 | 34.8 |
| Spain | 3 | 4.3 | Overall confidence[‡] | | |
| Australia | 1 | 1.4 | High | 0 | 0.0 |
| Turkey | 12 | 17.4 | Moderate | 1 | 1.4 |
| Iran | 2 | 2.9 | Low | 6 | 8.7 |
| China (Taiwan included) | 13 | 18.8 | Critically low | 62 | 89.9 |
| India | 3 | 4.3 | Risk of bias assessment | | |
| The Philippines | 1 | 1.4 | Lower ($\leq$2 critical flaws) | 21 | 30.4 |
| | | | Higher (>2 critical flaws) | 48 | 69.6 |

[‡]*Overall confidence* in the results of the review rated as: *High*: No or one non-critical weakness: the systematic review provides an accurate and comprehensive summary of the results of the available studies that address the question of interest; *Moderate*: More than one non-critical weakness: the systematic review has more than one weakness but no critical flaws. It may provide an accurate summary of the results of the available studies that were included in the review; *Low*: One critical flaw with or without non-critical weaknesses: the review has a critical flaw and may not provide an accurate and comprehensive summary of the available studies that address the question of interest; and *Critically low*: More than one critical flaw with or without non-critical weaknesses: the review has more than one critical flaw and should not be relied on to provide an accurate and comprehensive summary of the available studies (Shea et al., 2017).

than one flaw in any of the critical domains (see Table 1) and, when dichotomized, almost 70% ($n = 49$) had a high RoB with more than two critical flaws.

### 3.3. Identification of factors associated to risk of bias (RoB)

Several factors were significantly associated to low RoB assessment (see Table 2). The adherence to PRISMA guidelines increased the probability of a lower RoB (OR = 5.5; 95%CI: 1.8–16.6), as well as the most recent studies (OR = 7.4; 95%CI: 1.5–35.3), a higher number of authors (OR = 1.4; 95%CI: 1.1–1.9), when the corresponding author was from any European country (OR = 3.7; 95%CI: 1.1–12.8), and when the study was published in the health educational area compared to the educational area of knowledge (OR = 13.4; 95%CI: 3.6–49.6). All other were not associated with the RoB level.

## 4. Discussion

In our study, the overall confidence in the results, using the AMSTAR 2, was found to be critically low in most of the included MAs of experimental educational interventions designed to improve academic performance. This means that there was more than one critical flaw in the dimensions measured that could affect the overall credibility of the published MAs. To the best of our knowledge, this is the first methodological systematic review in Education using a structured instrument specifically developed to identify critical weaknesses affecting the confidence in the results of a systematic review, AMSTAR 2. Two previous methodological reviews used different checklist specifically developed by their authors (Ahn et al., 2012; Eser & Yurtçu, 2020). With this limitation in mind, several weaknesses were found suggesting a low confidence. The first one addressed the validity of a total of 56 MA published in the 2000s and the authors concluded that there was a lack of reported information and a need to increase transparency in the use of methods used (Ahn et al., 2012). The second, though limited to MAs published in educational sciences

**Table 2.** Association of characteristics of included studies with the overall confidence based on the risk of bias (RoB).

| | Risk of bias[#] | | | | 95%CI | | |
|---|---|---|---|---|---|---|---|
| | Higher *n* (%) | Lower *n* (%) | *p*-Value | OR | Inf | Sup | *p*-Value |
| Total | 48 (69.6) | 21 (30.4) | | | | | |
| PRISMA adherence | | | | | | | |
| No | 37 (77.1) | 8 (38.1) | | Ref. | | | |
| Yes | 11 (22.9) | 13 (61.9) | **0.002** | 5.5 | 1.8 | 16.6 | **0.003** |
| Year of publication (mean, *SD*) | | | | | | | |
| 2009–2016 | 21 (43.8) | 2 (9.5) | | Ref. | | | |
| 2017–2022 | 27 (56.2) | 19 (90.5) | **0.006** | 7.4 | 1.5 | 35.3 | **0.012** |
| Impact factor (mean, *SD*) | 2.18 (1.9) | 1.51 (0.4) | 0.952 | 1.0 | 0.7 | 1.4 | 0.951 |
| Number of authors (mean, *SD*) | 3.1 (1.8) | 5.1 (3.6) | **0.023** | 1.4 | 1.1 | 1.9 | **0.015** |
| Search strategy with geographic restrictions | | | | | | | |
| National restrictions | 10 (20.8) | 3 (14.3) | | Ref. | | | |
| International | 38 (79.2) | 18 (85.7) | 0.740* | 1.6 | 0.4 | 6.4 | 0.523 |
| Country | | | | | | | |
| Asia and Australia | 24 (50.0) | 8 (38.1) | | Ref. | | | |
| Europe | 8 (16.7) | 10 (47.6) | | 3.7 | 1.1 | 12.8 | **0.035** |
| North-America | 16 (33.3) | 3 (14.3) | **0.021** | 0.6 | 0.1 | 2.4 | 0.563 |
| Area of education | | | | | | | |
| Education | 36 (75.0) | 7 (33.3) | | Ref. | | | |
| Psychology | 7 (14.6) | 1 (4.8) | | 0.7 | 0.1 | 6.9 | 0.735 |
| Health | 5 (10.2) | 13 (61.9) | **<0.001*** | 13.4 | 3.6 | 49.6 | **<0.001** |
| Type of intervention | | | | | | | |
| Methodological strategies | 24 (50.0) | 6 (28.6) | | Ref. | | | |
| Organizational strategies | 11 (22.9) | 6 (28.6) | | 2.2 | 0.6 | 8.3 | 0.253 |
| Physical activities and others | 5 (10.4) | 5 (23.8) | | 4.0 | 0.9 | 18.4 | 0.076 |
| Combination | 8 (16.7) | 4 (19.0) | 0.296* | 2.0 | 0.5 | 8.9 | 0.364 |
| Target population | | | | | | | |
| Primary, secondary school and disabilities | 8 (16.7) | 2 (10.0) | | Ref. | | | |
| Higher education | 8 (16.7) | 7 (35.0) | | 2.3 | 0.4 | 12.4 | 0.320 |
| All levels included | 16 (33.3) | 3 (15.0) | | 0.5 | 0.1 | 3.1 | 0.453 |
| Combination of levels | 16 (33.3) | 8 (40.0) | 0.222* | 1.3 | 0.3 | 6.4 | 0.720 |

Model 1: each row represents the association of each independent variable with the low/high risk of bias (dependent variables) in a bivariate logistic regression mode.
Bold Values showing significant differences.
*Fisher's exact test.
[#]Risk of bias: Lower (<2 critical flaws); Higher (>2 critical flaws).

between 2010 and 2019 in Turkey, also detected several flaws that need to be improved (e.g. not establishing a previous hypothesis, not using a flow diagram, and some statistical deficiencies). Besides, this lack of confidence seems to be more disseminated than expected and not limited to Education. Several other systematic reviews of the confidence of MAs have been published with similar results (Bidhendi Yarandi et al., 2021; Cafri et al., 2010; Delaney et al., 2005; Hameed et al., 2020; Huedo-Medina et al., 2016; López-Nicolás et al., 2022). However, others have found an overall good quality of reports of MAs (Panahi et al., 2020; Sharma & Oremus, 2018; Suebnukarn et al., 2010).

Other instruments have been developed to assess RoB in systematic reviews in addition to AMSTAR 2. For example, ROBIS (Risk Of Bias In Systematic reviews) was designed as another assessment tool to evaluate the level of bias present in a systematic review and both instruments had similar measurement properties (Pieper et al., 2019; Swierz et al., 2021). However, AMSTAR 2 provides a broader RoB assessment than ROBIS (Shea et al., 2017). The original AMSTAR was design and validated as a practical critical appraisal tool for professionals to evaluate the quality of systematic reviews of randomized controlled trials of interventions (Shea, Bouter, et al., 2007; Shea, Grimshaw, et al., 2007). This instrument was updated to include non-randomized studies and changes were introduced to rate the overall confidence in the results (Shea et al., 2017). Briefly, instead of combining items to create an overall score (Jüni et al., 1999), authors recommended to identified critical domains and evaluate the overall confidence by a combination of critical and non-critical items in high, moderate, low, and critically low. The construction of the instrument was principally based on the original validated AMSTAR, but it also relied on the consensus of the expert panel and the feedback of other experts (Shea et al., 2017). AMSTAR 2 identifies a list of seven critical domains and the rating overall confidence in the results of the review rely mainly in those weaknesses considered as critical so that they should reduce the confidence in the findings of the review. However, their authors recognized that depending on different circumstances the

consideration of critical of some domains might be arguable or might change over time. For example, the most non-compliance critical item in our study was related to the presence of a list of excluded studies and the reason for exclusion provided by review authors. Though important in terms of improving transparency of all stages of conducting a MAs, it is not clear whether the absence of the excluded list might impact the results so high to be considered as critical. In the first version of the PRISMA statement (Moher et al., 2009) the suggestion was to 'give numbers of studies assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram'. It was not mandatory to report the reason of exclusion of each of the excluded studies individually. However, in the latest up-dated version, PRISMA 2020 statement, this item was explicitly included ('16b Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded') (Page, McKenzie, et al., 2021). As original authors explicitly suggest, the proposed list of critical domains is a suggestion that can be modified by users (Shea, Bouter, et al., 2007). However, we decided not to change the originally proposed seven critical domains but to modify the evaluation system using a broader and less restrictive one based on a categorization (lower vs. higher RoB) according to the number of critical flaws.

The second goal of this study was to assess the RoB related to whether authors mentioned the use of the PRISMA guideline, a checklist specifically design to improve the quality of the reporting of systematic reviews and MAs (Moher et al., 2009; Page, McKenzie, et al., 2021). Our results suggest that using this checklist is associated with a lower RoB due to a higher methodological quality. Though different constructs, it seems reasonable that methodological and reporting quality of MAs are closely correlated. In general, this result is in accordance with other studies that evaluated the effect of the endorsement of the PRISMA guideline when writing on the methodological quality (Cullis et al., 2017; Panic et al., 2013; Sharma & Oremus, 2018; Yuan et al., 2021).

Our third goal was to explore potential predictors of a low RoB and several factors were significantly associated to a lower RoB assessment. Those MAs more recently published (after 2017) had a significant lower RoB than those published previously. This finding might suggest an improvement in the methodological quality over time (De Vito et al., 2007; Delaney et al., 2005; Panahi et al., 2020). This pattern might reflect the increasing influence of the different instruments developed to improve the quality of reporting [e.g. QUORUM (Moher et al., 1999), PRISMA (Moher et al., 2009) and PRISMA 2020 (Page, McKenzie, et al., 2021)], and design [AMSTAR (Shea, Grimshaw, et al., 2007) and its update, AMSTAR 2 (Shea et al., 2017)], among others. If this is true, the generalization of these instruments with initiatives, such as the EQUATOR Network (https://www.equator-network.org/) (Simera et al., 2008) or the commitment to use checklists, such as the PRISMA statement, by high-impact journals [e.g. PLOS one (https://journals.plos.org/plosone/s/submission-guidelines#loc-guidelines-for-specific-study-types)] will contribute to the necessary improvement of the quality of future MAs.

First author location was another factor significantly associated to RoB. Those from Europe were at lower Rob compared to those from Asia and Australia, but those from North America did not differ from the latest. This result should be interpreted and generalized with caution as it is centered on MAs of experimental designs in Education and might not apply to other fields of knowledge. Our results support previous findings of a improvable quality of MAs from Turkey (Eser & Yurtçu, 2020). Contradictory results have been published in other areas of knowledge. For example, a lower reporting quality of North American first authors in diagnostic pathology (Liu et al., 2017), but no significant association was found in vaccinology (De Vito et al., 2007).

Education was divided in three main broad areas according to the context where the experimental intervention was performed, that is, education itself, and education in psychology, and in other health areas of education. As expected, the area with a lower RoB was Health Education compared to Education. This finding is not surprising as the design of systematic reviews and MA has drastically improved over the past decades in the health sciences (Gould, 2016; Moher et al., 1999, 2009; Page, McKenzie, et al., 2021; Shea, Grimshaw, et al., 2007; Shea et al., 2017) while it is being applied more recently in Education (Higgins, 2016; Polanin et al., 2017). Those MAs with lower RoB had a significantly higher mean number of authors. This result might be associated to the area of education as, in general, studies performed in the health area have a higher number of authors than those performed and published in Education. We performed further analysis to confirm this explanation and we found significant differences among areas [Health: mean = 5.3 ($SD = 3.4$); Education 3.1 (2.1) and Psychology 3.25 (1.7),

*p*-value = 0.010], specifically among Health and Education in the *post-hoc* analyses ($p = 0.008$), but not among Education and Psychology ($p = 0.993$) or Health and Psychology ($p = 0.136$).

### *4.1. Strengths and limitations*

Among the strengths of this review, it is worth highlighting that the protocol was previously registered in an international platform, a systematic review was performed and two researchers independently participated in the selection and data extraction process resolving disagreements by consensus or by the intervention of a third researcher. Finally, to the best of our knowledge, the present study is the first to comprehensively assess the RoB in Education based on a structured tool, the AMSTAR 2 appraisal tool. However, there are several limitations to this study. First, we defined a restrictive inclusion criterion focused only on educational interventions with an experimental design and any academic outcome. This restriction limits the external validity of our results to other SRs/MAs of other topics in Education. Future RoB assessment of SRs/MAs focused on other outcomes or designs will give us a more accurate analysis of the problem. Second, although our systematic literature search was performed in four independent databases, no grey literature was searched. However, this might only have served to identify more MA with a higher RoB and lower methodological quality. Finally, the identified predictive factors should be interpreted with caution as the RoB evaluation was based on a rational, but not validated, categorization criteria based on the number of critical flaws. Other classification criteria could lead to different results.

## 5. Conclusions

Our study raises concerns regarding the methodological quality of published MAs in Education focused on academic results as the outcome of interest. Some implications can be addressed from our results. First, published SRs and MAs are not free of potential methodological flaws that might reduce the confidence in their results. Researchers and editors should be aware of this situation, but also potential readers and related stakeholders in education. The value of an SR or an MA is not only as good as the evidence of the included studies but also as the quality of the methods used in its own design. If MAs are considered as the highest level in the hierarchy of evidence (Guyatt et al., 2000), or, at least, as the lens through other types of studies (Murad et al., 2016), the methods and reporting quality of this studies must be at the highest level. Several tools to help researchers design and carry out, such as the PRISMA 2020 statement and AMSTAR 2, have been developed in recent years and their compliance should be improved. Finally, our findings suggest a call for more research and education on the methodology of MAs may improve the design and reporting transparency and quality of future MAs in Education.

## Ethical approval

This study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Ethical Committee for Psycho-educational Research of the University of Granada (201-300 Academic Ranking of World Universities, Shanghai) (n°: 1858/CEIH/2020).

## Author contributions

Micaela Sánchez-Martín (MSM), Marta Gutiérrez-Sánchez (MGS), Eva María Olmedo-Moreno (EOM), and Fernando Navarro-Mateu (FNM) designed the study. MSM and MGS participated in the selection process and independently extracted the data. EOM and FNM participated as third evaluators when necessary. All authors participated in the interpretation of the results. MSM and FNM drafted the manuscript, and all authors critically revised it and approved the final version to be published.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

Research project: Transformation of learning in hybrid contexts for the educational and labor inclusion of especially vulnerable sectors with emphasis on UFM. TYNDALL/UFM. Reference: PID2020-119194RB-I00, financed by MCIN/AEI/10.13039/501100011033.

## About the authors

***Micaela Sánchez-Martín***, PhD, Associate Professor of Methodology Research and Diagnosis in Education at the University of Murcia (Spain).


***Marta Gutiérrez-Sánchez***, PhD, Associate Professor of Pedagogy at the University of Murcia (Spain).

***Eva María Olmedo-Moreno***, PhD, Professor of Methodology Research and Diagnosis in Education i at the University of Granada (Spain). She is Principal Research of Project R + D+I (Reference: PID2020-119194RB-I00, financed by MCIN/AEI/10.13039/501100011033) that support this study.

***Fernando Navarro-Mateu***, MD, PhD, Head of the Mental Health Teaching Unit and Adjunct Professor of Methodology in Psychology at the University of Murcia (Spain). His research interests focus on teaching evidence based practice (EBP), research methodology and the epidemiology of mental disorders.

## ORCID

Micaela Sánchez-Martín http://orcid.org/0000-0002-9107-574X
Marta Gutiérrez-Sánchez http://orcid.org/0000-0001-7302-9283
Eva María Olmedo-Moreno http://orcid.org/0000-0003-0558-1513
Fernando Navarro-Mateu http://orcid.org/0000-0002-3228-623X

## Data availability statement

The protocol was previously registered in OSF (Open Science Framework: https://doi.org/10.17605/OSF.IO/PSKN6). All data used has been included in the paper.

## References

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

Anguera, M. T. (2023). Revisitando las revisiones sistemáticas desde la perspectiva metodológica. *RELIEVE-Revista Electrónica De Investigación Y Evaluación Educativa*, *29*(1), art. M4. https://doi.org/10.30827/relieve.v29i1.27758

Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: How will we ever keep up? *PLOS Medicine*, *7*(9), e1000326. https://doi.org/10.1371/journal.pmed.1000326

Bidhendi Yarandi, R., Amiri, M., Ramezani Tehrani, F., & Behboudi-Gandevani, S. (2021). Effectiveness of antidiabetic agents for treatment of gestational diabetes: A methodological quality assessment of meta-analyses and network meta-analysis. *Journal of Diabetes Investigation*, *12*(12), 2247–2258. https://doi.org/10.1111/jdi.13603

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

Cullis, P. S., Gudlaugsdottir, K., & Andrews, J. (2017). A systematic review of the quality of conduct and reporting of systematic reviews and meta-analyses in paediatric surgery. *PLOS One*, *12*(4), e0175213. https://doi.org/10.1371/journal.pone.0175213

De Vito, C., Manzoli, L., Marzuillo, C., Anastasi, D., Boccia, A., & Villari, P. (2007). A systematic review evaluating the potential for bias and the methodological quality of meta-analyses in vaccinology. *Vaccine*, *25*(52), 8794–8806. https://doi.org/10.1016/j.vaccine.2007.10.034

Delaney, A., Bagshaw, S. M., Ferland, A., Manns, B., Laupland, K. B., & Doig, C. J. (2005). A systematic evaluation of the quality of meta-analyses in the critical care literature. *Critical Care*, *9*(5), R575–582. https://doi.org/10.1186/cc3803

Eser, M. T., & Yurtçu, M. (2020). A review of meta-analysis articles in educational sciences conducted between 2010 and 2019 in Turkey. *International Journal of Progressive Education*, *16*(6), 15–32. https://doi.org/10.29329/ijpe.2020.280.2

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

Gould, K. A. (2016). The EQUATOR network: A resource for authors. *Dimensions of Critical Care Nursing*, *35*(6), 350–350. https://doi.org/10.1097/DCC.0000000000000213

Gurevitch, J., Koricheva, J., Nakagawa, S., & Stewart, G. (2018). Meta-analysis and the science of research synthesis. *Nature*, *555*(7695), 175–182. https://doi.org/10.1038/nature25753

Guyatt, G. H., Haynes, R. B., Jaeschke, R. Z., Cook, D. J., Green, L., Naylor, C. D., Wilson, M. C., & Richardson, W. S. (2000). Users' guides to the medical literature: XXV. Evidence-based medicine: Principles for applying the users' guides to patient care. Evidence-Based Medicine Working Group. *JAMA*, *284*(10), 1290–1296. https://doi.org/10.1001/jama.284.10.1290

Hameed, I., Demetres, M., Tam, D. Y., Rahouma, M., Khan, F. M., Wright, D. N., Mages, K., DeRosa, A. P., Baltich Nelson, B., Pain, K., Delgado, D., Girardi, L. N., Fremes, S. E., & Gaudino, M. (2020). An assessment of the quality of current clinical meta-analyses. *BMC Medical Research Methodology*, *20*(1), 105. https://doi.org/10.1186/s12874-020-00999-9

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M., & Welch, V. (2022). *Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022)* (2nd ed.). Cochrane. www.training.cochrane.org/handbook

Higgins, S. (2016). Meta-synthesis and comparative meta-analysis of education research findings: Some risks and benefits. *Review of Education*, *4*(1), 31–53. https://doi.org/10.1002/rev3.3067

Holmqvist, M., & Ekström, M. L. (2024). A systematic review of research on educational superintendets. *Cugent Education*, *11*(1), 2307142. https://doi.org/10.1080/233118X.2024.2307142

Huedo-Medina, T. B., Garcia, M., Bihuniak, J. D., Kenny, A., & Kerstetter, J. (2016). Methodologic quality of meta-analyses and systematic reviews on the Mediterranean diet and cardiovascular disease outcomes: A review. *The American Journal of Clinical Nutrition*, *103*(3), 841–850. https://doi.org/10.3945/ajcn.115.112771

Hume, S., Brown, S. R., & Mahtani, K. R. (2023). School closures during COVID-19: An overview of systematic reviews. *BMJ Evidence-Based Medicine*, *28*(3), 164–174. https://doi.org/10.1136/bmjebm-2022-112085

Jin, S., Park, S.-M., Choi, S.-Y., Park, S. Y., & Kim, J.-H. (2023). Quality assessment of systematic reviews with meta-analysis in undergraduate nursing education. *Nurse Education Today*, *126*, 105833. https://doi.org/10.1016/j.nedt.2023.105833

Jüni, P., Witschi, A., Bloch, R., & Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*, *282*(11), 1054–1060. https://doi.org/10.1001/jama.282.11.1054

Liu, X., Kinzler, M., Yuan, J., He, G., & Zhang, L. (2017). Low reporting quality of the meta-analyses in diagnostic pathology. *Archives of Pathology & Laboratory Medicine*, *141*(3), 423–430. https://doi.org/10.5858/arpa.2016-0144-OA

López-López, J. A., Rubio-Aparicio, M., & Sánchez-Meca, J. (2022). Overviews of reviews: Concept and development. *Psicothema*, *34*(2), 175–181. https://doi.org/10.7334/psicothema2021.586

López-Nicolás, R., López-López, J. A., Rubio-Aparicio, M., & Sánchez-Meca, J. (2022). A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000–2020). *Behavior Research Methods*, *54*(1), 334–349. https://doi.org/10.3758/s13428-021-01644-z

Martinez-Calderon, J., Ho, E. K.-Y., Ferreira, P. H., Garcia-Muñoz, C., Villar-Alises, O., & Matias-Soto, J. (2023). A call for improving research on pain neuroscience education and chronic pain: An overview of systematic reviews. *The Journal of Orthopaedic and Sports Physical Therapy*, *53*(6), 353–368. https://doi.org/10.2519/jospt.2023.11833

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. QUality Of Reporting Of Meta-analyses. *Lancet*, *354*(9193), 1896–1900. https://doi.org/10.1016/s0140-6736(99)04149-5

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moreno, E. M., Montilla-Arechabala, C., & Maldonado, M. A. (2022). Effectiveness and characteristics of programs for developing oral competencies at university: A systematic review. *Cogent Education*, *9*(1), 2149224. https://doi.org/10.1080/2331186X.2022.2149224

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *Evidence-Based Medicine*, *21*(4), 125–127. https://doi.org/10.1136/ebmed-2016-110401

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*, n71. https://doi.org/10.1136/bmj.n71

Page, M. J., Moher, D., 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., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

Panahi, M. H., Mohseni, M., Bidhendi Yarandi, R., & Ramezani Tehrani, F. (2020). A methodological quality assessment of systematic reviews and meta-analyses of antidepressants effect on low back pain using updated AMSTAR. *BMC Medical Research Methodology*, *20*(1), 14. https://doi.org/10.1186/s12874-020-0903-9

Panic, N., Leoncini, E., de Belvis, G., Ricciardi, W., & Boccia, S. (2013). Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. *PLOS One*, *8*(12), e83138. https://doi.org/10.1371/journal.pone.0083138

Pieper, D., Puljak, L., González-Lorenzo, M., & Minozzi, S. (2019). Minor differences were found between AMSTAR 2 and ROBIS in the assessment of systematic reviews including both randomized and nonrandomized studies. *Journal of Clinical Epidemiology*, *108*, 26–33. https://doi.org/10.1016/j.jclinepi.2018.12.004

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

Sharma, S., & Oremus, M. (2018). PRISMA and AMSTAR show systematic reviews on health literacy and cancer screening are of good quality. *Journal of Clinical Epidemiology*, *99*, 123–131. https://doi.org/10.1016/j.jclinepi.2018.03.012

Shea, B. J., Bouter, L. M., Peterson, J., Boers, M., Andersson, N., Ortiz, Z., Ramsay, T., Bai, A., Shukla, V. K., & Grimshaw, J. M. (2007). External validation of a measurement tool to assess systematic reviews (AMSTAR). *PLOS One*, *2*(12), e1350. https://doi.org/10.1371/journal.pone.0001350

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., Tugwell, P., Moher, D., & Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*(1), 10. https://doi.org/10.1186/1471-2288-7-10

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

Simera, I., Altman, D. G., Moher, D., Schulz, K. F., & Hoey, J. (2008). Guidelines for reporting health research: The EQUATOR network's survey of guideline authors. *PLOS Medicine*, *5*(6), e139. https://doi.org/10.1371/journal.pmed.0050139

Suebnukarn, S., Ngamboonsirisingh, S., & Rattanabanlang, A. (2010). A systematic evaluation of the quality of meta-analyses in endodontics. *Journal of Endodontics*, *36*(4), 602–608. https://doi.org/10.1016/j.joen.2009.12.019

Swierz, M. J., Storman, D., Zajac, J., Koperny, M., Weglarz, P., Staskiewicz, W., Gorecka, M., Skuza, A., Wach, A., Kaluzinska, K., Bochenek-Cibor, J., Johnston, B. C., & Bala, M. M. (2021). Similarities, reliability and gaps in assessing the quality of conduct of systematic reviews using AMSTAR-2 and ROBIS: Systematic survey of nutrition reviews. *BMC Medical Research Methodology*, *21*(1), 261. https://doi.org/10.1186/s12874-021-01457-w

Yuan, M., Wu, J., Austin, R. E., Lista, F., & Ahmad, J. (2021). Evaluating the quality of systematic reviews and meta-analyses about breast augmentation using AMSTAR. *Aesthetic Surgery Journal. Open Forum*, *3*(3), ojab020. https://doi.org/10.1093/asjof/ojab020

## *References of included studies*

Costa, J. M., Lobato, G., & Melo, M. (2021). Four-component instructional design (4C/ID) model a meta-analysis on use and effect. *Learning Environments Research*, *25*(2), 445–463. https://doi.org/10.1007/s10984-021-09373-y

Ergen, B., & Kanadli, S. (2017). The effect of self-regulated learning strategies on academic achievement: A meta-analysis study. *Eurasian Journal of Educational Research*, *17*(69), 55–74. https://doi.org/10.14689/ejer.2017.69.4

Faramarzi, S., Shamsi, A., Samadi, M., & Ahmadzade, M. (2015). Meta-analysis of the efficacy of psychological and educational interventions to improve academic performance of students with learning disabilities in Iran. *Journal of Education and Health Promotion*, *4*(1), 58. https://doi.org/10.4103/2277-9531.162372

García-Hermoso, A., Ramírez-Vélez, R., Lubans, D. R., & Izquierdo, M. (2021). Effects of physical education interventions on cognition and academic performance outcomes in children and adolescents: A systematic review and meta-analysis. *British Journal of Sports Medicine*, *55*(21), 1224–1232. https://doi.org/10.1136/bjsports-2021-104112

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Sage Publications.

Hew, K. F., & Lo, C. K. (2018). Flipped classroom improves student learning in health professions education: A meta-analysis. *BMC Medical Education*, *18*(1), 38. https://doi.org/10.1186/s12909-018-1144-z

Hu, R., Gao, H., Ye, Y., Ni, Z., Jiang, N., & Jiang, X. (2018). Effectiveness of flipped classrooms in Chinese baccalaureate nursing education: A meta-analysis of randomized controlled trials. *International Journal of Nursing Studies*, *79*, 94–103. https://doi.org/10.1016/j.ijnurstu.2017.11.012

Hu, Y., Chen, C.-H., & Su, C.-Y. (2020). Exploring the effectiveness and moderators of block-based visual programming on student learning: A meta-analysis. *Journal of Educational Computing Research*, *58*(8), 1467–1493. https://doi.org/10.1177/0735633120945935

Jacobse, A. E., & Harskamp, E. G. (2011). *A meta-analysis of the effects of instructional interventions on students' mathematics achievement* (pp. 1–37). GION, Gronings Instituut Voor Onderzoek Van Onderwijs, Opvoeding En Ontwikkeling, Rijksuniversiteit Groningen. https://pure.rug.nl/ws/portalfiles/portal/10463070/A_Meta-Analysis_of_the_Effects_1.pdf

Kaçar, T., Terzi, R., Arıkan, İ., & Kırıkçı, A. C. (2021). The effect of inquiry-based learning on academic success: A meta-analysis study. *International Journal of Education and Literacy Studies*, *9*(2), 15. https://doi.org/10.7575/aiac.ijels.v.9n.2p.15

Kalaian, S. A., & Kasim, R. M. (2017). Effectiveness of various innovative learning methods in health science classrooms: A meta-analysis. *Advances in Health Sciences Education: Theory and Practice*, *22*(5), 1151–1167. https://doi.org/10.1007/s10459-017-9753-6

Karagöl, İ., & Esen, E. (2018). The effect of flipped learning approach on academic achievement: A meta-analysis study. *Hacettepe University Journal of Education, 34*(3), 1–20. https://doi.org/10.16986/HUJE.2018046755

Karich, A. C., Burns, M. K., & Maki, K. E. (2014). Updated meta-analysis of learner control within educational technology. *Review of Educational Research*, *84*(3), 392–410. https://doi.org/10.3102/0034654314526064

Kim, J., Gilbert, J., Yu, Q., & Gale, C. (2021). Measures matter: A meta-analysis of the effects of educational apps on preschool to grade 3 children's literacy and math skills. *AERA Open*, *7*, 233285842110041. https://doi.org/10.1177/23328584211004183

King-Sears, M. E., Stefanidis, A., Berkeley, S., & Strogilos, V. (2021). Does co-teaching improve academic achievement for students with disabilities? A meta-analysis. *Educational Research Review*, *34*, 100405. https://doi.org/10.1016/j.edurev.2021.100405

Korpershoek, H., Harms, T., De Boer, H., Van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students' academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, *86*(3), 643–680. https://doi.org/10.3102/0034654315626799

Låg, T., & Sæle, R. G. (2019). Does the flipped classroom improve student learning and satisfaction? A systematic review and meta-analysis. *AERA Open*, *5*(3), 233285841987048. https://doi.org/10.1177/2332858419870489

Leung, K. C. (2015). Preliminary empirical model of crucial determinants of best practice for peer tutoring on academic achievement. *Journal of Educational Psychology*, *107*(2), 558–579. https://doi.org/10.1037/a0037698

Linden, M. A., Glang, A. E., & McKinlay, A. (2018). A systematic review and meta-analysis of educational interventions for children and adolescents with acquired brain injury. *NeuroRehabilitation*, *42*(3), 311–323. https://doi.org/10.3233/NRE-172357

May, R. J., Tyndall, I., McTiernan, A., Roderique-Davies, G., & McLoughlin, S. (2021). The impact of the SMART program on cognitive and academic skills: A systematic review and meta-analysis. *British Journal of Educational Technology*, *53*(5), 1244–1261. https://doi.org/10.1111/bjet.13192

Moore, D. A., Russell, A. E., Matthews, J., Ford, T. J., Rogers, M., Ukoumunne, O. C., Kneale, D., Thompson-Coon, J., Sutcliffe, K., Nunns, M., Shaw, L., & Gwernan-Jones, R. (2018). School-based interventions for attention-deficit/hyperactivity disorder: A systematic review with multiple synthesis methods. *Review of Education*, *6*(3), 209–263. https://doi.org/10.1002/rev3.3149

Norris, E., Van Steen, T., Direito, A., & Stamatakis, E. (2019). Physically active lessons in schools and their impact on physical activity, educational, health and cognition outcomes: A systematic review and meta-analysis. *British Journal of Sports Medicine*, *54*(14), 826–838. https://doi.org/10.1136/bjsports-2018-100502

Oh-Young, C., & Filler, J. (2015). A meta-analysis of the effects of placement on academic and social skill outcome measures of students with disabilities. *Research in Developmental Disabilities*, *47*, 80–92. https://doi.org/10.1016/j.ridd.2015.08.014

Orhan, A. (2019). The effect of flipped learning on students' academic achievement: A meta-analysis study. *Çukurova Üniversitesi Eğitim Fakültesi Dergisi*, *48*(1), 368–396. http://dergipark.gov.tr/cuefd

Petersen-Brown, S. M., Henze, E. E. C., Klingbeil, D. A., Reynolds, J. L., Weber, R. C., & Codding, R. S. (2019). The use of touch devices for enhancing academic achievement: A meta-analysis. *Psychology in the Schools*, *56*(7), 1187–1206. https://doi.org/10.1002/pits.22225

Phelps, R. P. (2019). Test frequency, stakes, and feedback in student achievement: A meta-analysis. *Evaluation Review*, *43*(3–4), 111–151. https://doi.org/10.1177/0193841X19865628

Robbins, S. B., Oh, I.-S., Le, H., & Button, C. (2009). Intervention effects on college performance and retention as mediated by motivational, emotional, and social control factors: Integrated meta-analytic path analyses. *The Journal of Applied Psychology*, *94*(5), 1163–1184. https://doi.org/10.1037/a0015738

Saw, K. N. N., & Han, B. (2021). Effectiveness of successful intelligence training program: A meta-analysis. *PsyCh Journal*, *10*(3), 323–339. https://doi.org/10.1002/pchj.443

Sayyah, M., Shirbandi, K., Saki-Malehi, A., & Rahim, F. (2017). Use of a problem-based learning teaching model for undergraduate medical and nursing education: A systematic review and meta-analysis. *Advances in Medical Education and Practice*, *8*, 691–700. https://doi.org/10.2147/AMEP.S143694

Semerci, Ç., & Batdi, V. (2015). A meta-analysis of constructivist learning approach on learners' academic achievements, retention and attitudes. *Journal of Education and Training Studies*, *3*(2), 171–180. https://doi.org/10.11114/jets.v3i2.644

Sneck, S., Viholainen, H., Syväoja, H., Kankaapää, A., Hakonen, H., Poikkeus, A.-M., & Tammelin, T. (2019). Effects of school-based physical activity on mathematics performance in children: A systematic review. *The International Journal of Behavioral Nutrition and Physical Activity*, *16*(1), 109. https://doi.org/10.1186/s12966-019-0866-6

Steele, J. L., Bozick, R., & Davis, L. M. (2016). Education for incarcerated juveniles: A meta-analysis. *Journal of Education for Students Placed at Risk*, *21*(2), 65–89. https://doi.org/10.1080/10824669.2015.1133308

Sugano, S. G., & Nabua, E. B. (2020). Meta-analysis on the effects of teaching methods on academic performance in chemistry. *International Journal of Instruction*, *13*(2), 881–894. https://doi.org/10.29333/iji.2020.13259a

Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, *94*, 252–275. https://doi.org/10.1016/j.compedu.2015.11.008

Tan, C., Yue, W.-G., & Fu, Y. (2017). Effectiveness of flipped classrooms in nursing education: Systematic review and meta-analysis. *Chinese Nursing Research*, *4*(4), 192–200. https://doi.org/10.1016/j.cnre.2017.10.006

Tokac, U., Novak, E., & Thompson, C. G. (2019). Effects of game-based learning on students' mathematics achievement: A meta-analysis. *Journal of Computer Assisted Learning*, *35*(3), 407–420. https://doi.org/10.1111/jcal.12347

Warren, J. L. (2012). Does service-learning increase student learning? A meta-analysis. *Michigan Journal of Community Service Learning*, *18*(2), 56–61.

Watson, A., Timperio, A., Brown, H., Best, K., & Hesketh, K. D. (2017). Effect of classroom-based physical activity interventions on academic and physical activity outcomes: A systematic review and meta-analysis. *The International Journal of Behavioral Nutrition and Physical Activity*, *14*(1), 114. https://doi.org/10.1186/s12966-017-0569-9

Wilson, A. B., Brown, K. M., Misch, J., Miller, C. H., Klein, B. A., Taylor, M. A., Goodwin, M., Boyle, E. K., Hoppe, C., & Lazarus, M. D. (2019). Breaking with tradition: A scoping meta-analysis analyzing the effects of student-centered learning and computer-aided instruction on student performance in anatomy: Effectiveness of student centered and computer-aided learning in anatomy. *Anatomical Sciences Education*, *12*(1), 61–73. https://doi.org/10.1002/ase.1789

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2020). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

Yorio, P. L., & Ye, F. (2012). A meta-analysis on the effects of service-learning on the social, personal, and cognitive outcomes of learning. *Academy of Management Learning & Education*, *11*(1), 9–27. https://doi.org/10.5465/amle.2010.0072

Zheng, B., Warschauer, M., Lin, C. H., & Chang, C. (2016). Learning in one-to-one laptop environments: A meta-analysis and research synthesis. *Review of Educational Research*, *86*(4), 1052–1084. https://doi.org/10.3102/0034654316628645

## Appendix A: Search strategy by data base

| Data base | Search strategy | Document numbers |
|---|---|---|
| PubMed | ((((("academe"[All Fields] OR "academic"[All Fields] OR "academic s"[All Fields] OR "academical"[All Fields] OR "academically"[All Fields] OR "academics"[All Fields]) AND ("outcome"[All Fields] OR "outcomes"[All Fields])) OR (("academe"[All Fields] OR "academic"[All Fields] OR "academic s"[All Fields] OR "academical"[All Fields] OR "academically"[All Fields] OR "academics"[All Fields]) AND ("result"[All Fields] OR "resultant"[All Fields] OR "resultants"[All Fields] OR "resulted"[All Fields] OR "resulting"[All Fields] OR "results"[All Fields])) OR ("academic success"[MeSH Terms] OR ("academic"[All Fields] AND "success"[All Fields]) OR "academic success"[All Fields] OR ("academic"[All Fields] AND "achievement"[All Fields]) OR "academic achievement"[All Fields]) OR ("academic performance"[MeSH Terms] OR ("academic"[All Fields] AND "performance"[All Fields]) OR "academic performance"[All Fields])) AND ((("educability"[All Fields] OR "educable"[All Fields] OR "educates"[All Fields] OR "education"[MeSH Subheading] OR "education"[All Fields] OR "educational status"[MeSH Terms] OR ("educational"[All Fields] AND "status"[All Fields]) OR "educational status"[All Fields] OR "education"[MeSH Terms] OR "education s"[All Fields] OR "educational"[All Fields] OR "educative"[All Fields] OR "educator"[All Fields] OR "educator s"[All Fields] OR "educators"[All Fields] OR "teaching"[MeSH Terms] OR "teaching"[All Fields] OR "educate"[All Fields] OR "educated"[All Fields] OR "educating"[All Fields] OR "educations"[All Fields]) AND ("intervention s"[All Fields] OR "interventions"[All Fields] OR "interventive"[All Fields] OR "methods"[MeSH Terms] OR "methods"[All Fields] OR "intervention"[All Fields] OR "interventional"[All Fields])) OR (("educability"[All Fields] OR "educable"[All Fields] OR "educates"[All Fields] OR "education"[MeSH Subheading] OR "education"[All Fields] OR "educational status"[MeSH Terms] OR ("educational"[All Fields] AND "status"[All Fields]) OR "educational status"[All Fields] OR "education"[MeSH Terms] OR "education s"[All Fields] OR "educational"[All Fields] OR "educative"[All Fields] OR "educator"[All Fields] OR "educator s"[All Fields] OR "educators"[All Fields] OR "teaching"[MeSH Terms] OR "teaching"[All Fields] OR "educate"[All Fields] OR "educated"[All Fields] OR "educating"[All Fields] OR "educations"[All Fields]) AND ("education"[MeSH Subheading] OR "education"[All Fields] OR "training"[All Fields] OR "education"[MeSH Terms] OR "train"[All Fields] OR "train s"[All Fields] OR "trained"[All Fields] OR "training s"[All Fields] OR "trainings"[All Fields] OR "trains"[All Fields])) OR (("educability"[All Fields] OR "educable"[All Fields] OR "educates"[All Fields] OR "education"[MeSH Subheading] OR "education"[All Fields] OR "educational status"[MeSH Terms] OR ("educational"[All Fields] AND "status"[All Fields]) OR "educational status"[All Fields] OR "education"[MeSH Terms] OR "education s"[All Fields] OR "educational"[All Fields] OR "educative"[All Fields] OR "educator"[All Fields] OR "educator s"[All Fields] OR "educators"[All Fields] OR "teaching"[MeSH Terms] OR "teaching"[All Fields] OR "educate"[All Fields] OR "educated"[All Fields] OR "educating"[All Fields] OR "educations"[All Fields]) AND ("program"[All Fields] OR "program s"[All Fields] OR "programe"[All Fields] OR "programed"[All Fields] OR "programes"[All Fields] OR "programing"[All Fields] OR "programmability"[All Fields] OR "programmable"[All Fields] OR "programmably"[All Fields] OR "programme"[All Fields] OR "programme s"[All Fields] OR "programmed"[All Fields] OR "programmer"[All Fields] OR "programmer s"[All Fields] OR "programmers"[All Fields] OR "programmes"[All Fields] OR "programming"[All Fields] OR "programmings"[All Fields] OR "programs"[All Fields]))))) AND ((meta-analysis[Filter]) AND (2009/1/1:2022/4/30[pdat]) AND (english[Filter] OR spanish[Filter]))<br>Filters: Meta-Analysis, English, Spanish, from 2009/1/1 to 2022/4/30 | 745 |
| ERIC | (metaanalysis or meta-analysis or meta analysis) AND (academic outcomes or academic achievement or academic performance or academic success or academic skills) AND (educational intervention or educational training or educational program)<br>Filters:<br>Limitadores:Fecha de publicación: 20090101-20220431<br>mpliadores: Aplicar materias equivalents<br>Especificar por Language: English<br>Especificar por Subject: meta análisis<br>Modos de búsqueda: Booleano/Frase | 458 |
| WOS | metaanalysis OR meta-analysis OR meta analysis (Topic) and academic outcomes OR academic achievement OR academic performance OR academic success OR academic skills (Topic) and educational intervention OR educational training OR educational program (Topic) and 2022 or 2021 or 2020 or 2019 or 2018 or 2017 or 2016 or 2015 or 2014 or 2013 or 2012 or 2011 or 2010 or 2009 (Publication Years) and Articles or Review Articles (Document Types)<br>Filters:<br>Publication years: 2009-2020<br>Document types: Articles or Review articles | 536 |
| SCOPUS | (TITLE-ABS-KEY (metanalysis OR meta-analysis OR meta AND analysis) AND TITLE-ABS-KEY (academic AND outcomes OR academic AND achievement OR academic AND performance OR academic AND success OR academic AND skills) AND TITLE-ABS-KEY (educational AND intervention OR educational AND training OR educational AND program))<br>Filters: Journal | 36 |

## Appendix B: List of excluded studies with the main reason for exclusion

| First author | Publication year | doi | Main reason for exclusion |
|---|---|---|---|
| Alvarez-Bueno | 2020 | 10.1080/02640414.2020.1720496 | No educational intervention |
| Barger | 2019 | 10.1037/bul0000201 | No educational intervention |
| Barranquero-Herbosa | 2022 | 10.1016/j.ijnurstu.2022.104327 | No meta-analysis |
| Baskin | 2010 | 10.1037/a0019652 | No educational intervention |
| Bedrow | 2021 | 10.3102/00346543211019122 | No full text |
| Beelmann | 2020 | 10.1002/ijop.12725 | No meta-analysis |
| Bilal | 2019 | 10.1016/j.sjbs.2017.10.024 | No educational intervention |
| Bolinskia | 2020 | 10.1016/j.invent.2020.100321 | No educational intervention |
| Bonastre | 2021 | 10.22550/REP79-2-2021-05 | No academic performance |
| Bowman-Perrott | 2015 | 10.1177/0145445514551383 | No educational intervention |
| Bradford | 2021 | 10.1187/cbe.20-03-0046 | No educational intervention |
| Braithwaite | 2016 | 10.1016/j.paid.2016.03.040 | No experimental design |
| Burns | 2016 | 10.1037/spq0000117 | No educational intervention |
| Camacho-Morles | 2021 | 10.1007/s10648-020-09585-3 | No educational intervention |
| Cawthon | 2013 | 10.1353/ aad.2013.0023 | Systematic review without metanalysis |
| Ciocanel | 2017 | 10.1007/s10964-016-0555-6 | No educational intervention |
| Conley | 2015 | 10.1007/s11121-015-0543-1 | No educational intervention |
| Conway | 2009 | 10.1080/00986280903172969 | No academic performance |
| Cooper | 2020 | 10.1111/1540-5834.00064 | No educational intervention |
| Curlette | 2014 | | Book Chapter |
| Dignath | 2008 | 10.1007/s11409-008-9029-x | Publication Date |
| Fedewa | 2021 | 10.1080/02701367.2011.10599785 | No academic performance |
| Gersten | 2020 | 10.1080/19345747.2019.1689591 | No academic performance |
| Haverkamp | 2020 | 10.1080/02640414.2020.1794763 | No academic performance |
| Hedberg | 2014 | 10.1177/0193841X14554212 erx.sagepub | No meta-analysis |
| Huang | 2019 | 10.1080/0142159X.2019.1623386 | No academic performance |
| Iachinia | 2015 | 10.1080/1754730X.2015.1044252 | No educational intervention |
| Jones | 2013 | 10.3109/0142159X.2013.806983 | No academic performance |
| Katesÿ | 2018 | 10.1016/j.compedu.2018.08.012 | No educational intervention |
| Ledbetter-Cho | 2018 | 10.1007/s1080 3-018-3573-2 | No control group |
| Lester | 2020 | 10.1007/s10964-019-01188-8 | No educational intervention |
| Liu | 2016 | 10.2196/jmir.4807 | No educational intervention |
| Lortie-Forgues | 2019 | 0.3102/0013189X19832850 | No educational intervention |
| Magnusona | 2016 | 10.1016/j.ecresq.2015.12.021 | No educational intervention |
| Martin | 2017 | 10.1080/02701367.2017.1294244 | No meta-analysis |
| Masini | 2019 | 10.1016/j.jsams.2019.10.008 | No academic performance |
| Namjin | 2015 | | No full text |
| Owen | 2022 | 10.1249/MSS.0000000000002786 | No educational intervention |
| Pandey | 2015 | 10.1001/jamapediatrics.2018.0232 | No academic performance |
| Rasberry | 2011 | 10.1016/j.ypmed.2011.01.027 | No meta-analysis |
| Reed | 2015 | 10.1007/s10648-014-9289-8 | No meta-analysis |
| Shin | 2019 | 10.1016/j.jsp.2019.03.005 | No academic performance |
| Szumski | 2017 | 10.1371/journal.pone.0270124 | No meta-analysis |
| Twomey | 2021 | 10.1177/1747021821993457 | No academic performance |

## Appendix C: Description of included documents in the systematic review

| First author (publication year) | Journal Impact factor | Explicit PRISMA adherence (Yes/No) | Number of authors | Geographic search restriction | Corresponding author's location | Area of education | Intervention analysed | Type of intervention | Population studied |
|---|---|---|---|---|---|---|---|---|---|
| Aydin (2021) | – | No | 4 | Regional or national | Asia and Australia | Education | Organizational | Flipped Classroom | Primary, secondary and higher education |
| Akar (2020) | 0.470 | No | 1 | Regional or national | Asia and Australia | Education | Methodological | Educational Technology | Primary, secondary and higher education |
| Aktamis (2016) | 0.250 | No | 3 | Regional or national | Asia and Australia | Education | Methodological | Situation-based methodological approaches | Secondary school |
| Alegre-Ansuategui (2018) | 0.325 | No | 4 | International | Europe | Education | Organizational | Organizational strategies | Todas |
| Alvarez-Bueno (2017) | 3.337 | Sí | 6 | International | Europe | Health | Physical activities and others | Physical Activity | Primary and secondary school |
| Aspiranti (2018) | 0.247 | No | 3 | International | North-America | Education | Methodological | Educational Technology | Disabilities |
| Ayaz (2015) | – | No | 2 | Regional or national | Asia and Australia | Education | Methodological | Learning theories | Todas |
| Bai (2020) | 3.277 | No | 3 | International | Asia and Australia | Education | Methodological | Gamification | Primary, secondary and higher education |
| Balakrishnan (2021) | 0.852 | Sí | 7 | International | Asia and Australia | Health | Physical activities and others | Others | Higher education |
| Bas (2016) | 0.261 | No | 1 | Regional or national | Asia and Australia | Education | Methodological | Learning theories | Todas |
| Bedard (2019) | 1.023 | No | 5 | International | North-America | Health | Physical activities and others | Physical Activity | Preschool, primary and secondary school |
| Brierly (2021) | | Sí | 3 | International | Europe | Health | Organizational | Organizational strategies | Higher education |
| Capar (2015) | 0.218 | No | 2 | International | Asia and Australia | Education | Organizational | Organizational strategies | Todas |
| Cartiff (2021) | 2.623 | No | 3 | International | North-America | Psychology | Methodological | Learning theories | Todas |
| Celio (2011) | 1.244 | No | 3 | International | North-America | Education | Methodological | Situation-based methodological approaches | Todas |
| Cen (2021) | 0.658 | Sí | 4 | International | Asia and Australia | Health | Methodological | Situation-based methodological approaches | Higher education |
| Chandran (2022) | 0.852 | Sí | 8 | International | Asia and Australia | Health | Methodological | Educational Technology | Higher education |
| Chauhan (2016) | 2.654 | Sí | 1 | International | Asia and Australia | Education | Methodological | Educational Technology | Primary School |
| Chen (2019) | 3.216 | No | 2 | International | Asia and Australia | Education | Methodological | Situation-based methodological approaches | Todas |
| Chen (2018) | 0.802 | Sí | 17 | Regional or national | Asia and Australia | Health | Organizational | Organizational strategies | Higher education |
| Cheng (2018) | 0.977 | Sí | 3 | International | North-America | Education | Organizational | Flipped Classroom | Todas |
| Cheung (2013) | 3.148 | No | 2 | International | Asia and Australia | Education | Methodological | Educational Technology | Todas |
| Costa et al. (2021) | 0.950 | No | 3 | International | Europe | Education | Methodological | Situation-based methodological approaches | Primary, secondary and higher education |

(*continued*)

Continued.

| First author (publication year) | Journal Impact factor | Explicit PRISMA adherence (Yes/No) | Number of authors | Geographic search restriction | Corresponding author's location | Area of education | Intervention analysed | Type of intervention | Population studied |
|---|---|---|---|---|---|---|---|---|---|
| Dagva (2015) | 5.658 | Sí | 2 | International | Asia and Australia | Education | Methodological | Situation-based methodological approaches | Primary, secondary and higher education |
| De Boer (2014) | 5.658 | Sí | 3 | International | Europe | Education | Combination | Several methodological approaches | Primary and secondary school |
| De Greeff (2017) | 1.714 | Sí | 5 | International | Europe | Health | Physical activities and others | Physical Activity | Primary School |
| Dietrichson (2017) | 3.719 | Sí | 4 | International | Europe | Education | Combination | Several methodological approaches | Primary and secondary school |
| Donker (2013) | 3.148 | No | 5 | International | Europe | Education | Combination | Several methodological approaches | Primary and secondary school |
| Double (2019) | 3.805 | Sí | 3 | International | Europe | Psychology | Organizational | Organizational strategies | Primary, secondary and higher education |
| Ergen and Kanadli (2017) | 0.184 | No | 2 | Regional or national | Asia and Australia | Education | Methodological | Learning theories | Primary, secondary and higher education |
| Faramarzi et al. (2015) | – | Sí | 4 | Regional or national | Asia and Australia | Education | Combination | Several methodological approaches | Disabilities |
| García-Hermoso et al. (2021) | 3.867 | Sí | 4 | International | Europe | Health | Physical activities and others | Physical Activity | Preschool, primary and secondary school |
| Hew and Lo (2018) | 0.802 | Sí | 2 | International | Asia and Australia | Health | Organizational | Flipped Classroom | Higher education |
| Hu et al. (2020) | 1.050 | Sí | 3 | International | Asia and Australia | Education | Methodological | Educational Technology | Primary, secondary and higher education |
| Hu et al. (2018) | 1.556 | Sí | 6 | Regional or national | Asia and Australia | Health | Organizational | Flipped Classroom | Higher education |
| Jacobse and Harskamp (2011) | – | No | 2 | International | Europe | Education | Combination | Several methodological approaches | Higher education |
| Kaçar et al. (2021) | – | No | 4 | Regional or national | Asia and Australia | Education | Methodological | Situation-based methodological approaches | Todas |
| Kalaian and Kasim (2017) | 1.640 | Sí | 2 | International | North-America | Health | Combination | Several methodological approaches | Higher education |
| Karagöl and Esen (2018) | | No | 2 | International | Asia and Australia | Education | Organizational | Flipped Classroom | Preschool, primary and secondary school |
| Karich et al. (2014) | 5.658 | Sí | 3 | International | North-America | Education | Methodological | Educational Technology | Secondary and higher education |
| Kim et al. (2021) | 0.856 | Sí | 4 | International | North-America | Education | Methodological | Educational Technology | Primary, secondary and higher education |
| King-Sears et al. (2021) | 3.067 | No | 4 | International | North-America | Education | Organizational | Organizational strategies | Disabilities |
| Korpershoek et al. (2016) | 3.853 | Sí | 5 | International | Europe | Education | Combination | Several methodological approaches | Preschool and primary |
| Låg and Sæle (2019) | – | Sí | 2 | International | Europe | Education | Organizational | Flipped Classroom | Todas |
| Leung (2015) | 2.814 | No | 1 | International | Asia and Australia | Psychology | Organizational | Organizational strategies | Todas |
| Linden et al. (2018) | 0.642 | Sí | 3 | International | Europe | Health | Combination | Several methodological approaches | Disabilities |

(*continued*)

Continued.

| First author (publication year) | Journal Impact factor | Explicit PRISMA adherence (Yes/No) | Number of authors | Geographic search restriction | Corresponding author's location | Area of education | Intervention analysed | Type of intervention | Population studied |
|---|---|---|---|---|---|---|---|---|---|
| May et al. (2021) | 1.870 | No | 5 | International | Europe | Education | Methodological | Educational Technology | Todas |
| Moore et al. (2018) | 5.509 | Sí | 12 | International | Europe | Education | Combination | Several methodological approaches | Preschool, primary and secondary school |
| Norris et al. (2019) | 3.712 | Sí | 4 | International | Europe | Health | Physical activities and others | Physical Activity | Preschool, primary and secondary school |
| Oh-Young and Filler (2015) | 0.996 | No | 2 | International | North-America | Psychology | Physical activities and others | Others | Disabilities |
| Orhan (2019) | – | No | 1 | Regional or national | Asia and Australia | Education | Organizational | Flipped Classroom | Secondary and higher education |
| Petersen-Brown et al. (2019) | 0.676 | No | 6 | International | North-America | Psychology | Methodological | Educational Technology | Preschool, primary and secondary school |
| Phelps (2019) | 0.839 | No | 1 | International | North-America | Education | Physical activities and others | Others | Todas |
| Robbins et al. (2009) | 5.723 | Sí | 4 | International | North-America | Psychology | Combination | Several methodological approaches | Higher education |
| Saw and Han (2021) | – | No | 2 | International | Asia and Australia | Psychology | Methodological | Learning theories | Todas |
| Sayyah et al. (2017) | – | Sí | 4 | Regional or national | Asia and Australia | Health | Methodological | Situation-based methodological approaches | Higher education |
| Semerci and Batdi (2015) | – | No | 2 | International | Asia and Australia | Education | Methodological | Learning theories | Todas |
| Sneck et al. (2019) | 2.637 | Sí | 7 | International | Europe | Health | Physical activities and others | Physical Activity | Preschool, primary and secondary school |
| Steele et al. (2016) | 0.729 | No | 3 | International | North-America | Education | Combination | Several methodological approaches | Others |
| Sugano and Nabua (2020) | 0.535 | No | 2 | Regional or national | Asia and Australia | Education | Combination | Several methodological approaches | Secondary school |
| Sung et al. (2016) | 2.654 | Sí | 3 | International | Asia and Australia | Education | Methodological | Educational Technology | Todas |
| Tan et al. (2017) | – | No | 3 | International | Asia and Australia | Health | Organizational | Flipped Classroom | Higher education |
| Tokac et al. (2019) | 1.540 | Sí | 1 | International | North-America | Education | Methodological | Gamification | Primary and secondary school |
| Warren (2012) | – | No | 5 | International | North-America | Education | Organizational | Organizational strategies | Higher education |
| Watson et al. (2017) | 2.626 | No | 10 | International | Asia and Australia | Health | Methodological | Educational Technology | Primary School |
| Wilson et al. (2019) | 0.820 | No | 5 | International | North-America | Education | Methodological | Educational Technology | Higher education |
| Yang et al. (2020) | 7.504 | Sí | 2 | International | Asia and Australia | Psychology | Physical activities and others | Others | Todas |
| Yorio and Ye (2012) | 2.378 | No | 4 | International | North-America | Education | Organizational | Organizational strategies | Higher education |
| Zheng et al. (2016) | 3.853 | Sí | | International | North-America | Education | Methodological | Educational Technology | Todas |

## Appendix D: Methodological quality of the included meta-analyses using the a measurement tool to assess systematic reviews (AMSTAR-2)

| First author (publication year) | AMSTAR-2 quality items | | | | | | | | | | | | | | | | Overall confidence (AMSTAR-2)‡ | Risk of bias (RoB)# |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2† | 3 | 4† | 5 | 6 | 7† | 8 | 9† | 10 | 11† | 12 | 13† | 14 | 15† | 16 | | |
| Adin (2021) | Y | N | N | PY | N | N | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Akar (2020) | Y | N | Y | PY | N | N | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Aktamis (2016) | Y | N | N | N | N | N | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Alegre-Ansuategui (2018) | Y | N | N | PY | N | N | N | Y | N | N | Y | N | N | Y | N | N | Critically low | High |
| Alvarez-Bueno (2017) | Y | Y | N | PY | Y | Y | N | Y | Y | N | Y | Y | Y | Y | Y | Y | Low | Low |
| Aspiranti (2018) | Y | N | N | PY | Y | Y | N | Y | N | N | N | N | N | Y | N | N | Critically low | High |
| Ayaz (2015) | Y | N | N | PY | N | Y | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Bai (2020) | Y | N | N | PY | Y | Y | N | N | Y | N | Y | Y | Y | Y | Y | N | Critically low | Low |
| Balakrishnan (2021) | Y | N | N | PY | Y | Y | PY | Y | Y | N | Y | Y | Y | Y | Y | N | Low | High |
| Bas (2016) | Y | N | N | N | N | Y | N | N | N | N | Y | Y | N | Y | N | N | Critically low | High |
| Bedard (2019) | Y | N | N | PY | Y | N | N | N | Y | N | Y | Y | Y | Y | Y | N | Critically low | Low |
| Brierly (2021) | Y | N | Y | PY | Y | Y | N | Y | Y | N | Y | Y | Y | Y | Y | N | Critically low | Low |
| C¸apar (2015) | Y | N | N | PY | N | N | N | PY | N | N | Y | N | N | Y | N | N | Critically low | High |
| Cartiff (2021) | Y | N | N | Y | N | Y | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Celio (2011) | Y | N | N | Y | N | N | N | N | N | N | N | N | N | N | N | N | Critically low | High |
| Cen (2021) | Y | N | Y | Y | Y | Y | N | PY | Y | N | Y | Y | N | Y | Y | Y | Critically low | High |
| Chandran (2022) | Y | Y | Y | Y | Y | Y | N | Y | Y | N | Y | Y | Y | Y | Y | Y | Low | Low |
| Chauhan (2016) | Y | N | N | PY | N | N | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Chen (2019) | Y | N | N | Y | N | Y | N | N | N | N | Y | N | N | Y | Y | Y | Critically low | High |
| Chen (2018) | Y | N | N | PY | N | Y | N | Y | Y | N | Y | Y | Y | Y | Y | Y | Critically low | Low |
| Cheng (2018) | Y | N | Y | PY | N | Y | N | N | N | N | Y | N | N | Y | Y | Y | Critically low | High |
| Cheung (2013) | Y | N | Y | PY | N | Y | N | Y | PY | N | Y | N | Y | Y | Y | N | Critically low | Low |
| Costa et al. (2021) | Y | N | N | PY | N | N | N | Y | N | N | Y | N | N | Y | N | N | Critically low | High |
| Dagva (2015) | Y | N | N | Y | N | Y | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| De Boer (2014) | Y | N | Y | PY | N | Y | N | Y | N | N | Y | N | N | Y | Y | N | Critically low | High |
| De Greeff (2017) | Y | N | N | Y | Y | N | Y | Y | N | N | Y | Y | Y | Y | Y | N | Critically low | Low |
| Dietrichson (2017) | Y | Y | Y | Y | N | N | N | PY | Y | N | Y | Y | Y | Y | N | N | Critically low | Low |
| Donker (2013) | Y | N | Y | PY | N | Y | N | Y | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Double (2019) | Y | N | Y | Y | Y | Y | N | PY | N | N | Y | Y | N | Y | Y | N | Critically low | High |
| Ergen and Kanadli (2017) | Y | N | N | PY | N | Y | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Faramarzi et al. (2015) | Y | N | N | N | N | Y | N | Y | N | N | N | N | N | N | N | N | Critically low | High |
| García-Hermoso et al. (2021) | Y | Y | N | Y | Y | Y | Y | Y | Y | N | Y | Y | Y | Y | Y | Y | Moderate | Low |
| Hew and Lo (2018) | Y | N | N | PY | N | Y | N | Y | Y | N | Y | Y | Y | Y | Y | Y | Critically low | Low |
| Hu et al. (2020) | Y | N | N | PY | Y | Y | N | Y | N | N | Y | N | N | Y | Y | Y | Critically low | High |
| Hu et al. (2018) | Y | N | N | PY | Y | Y | N | Y | Y | N | Y | N | Y | Y | Y | Y | Critically low | Low |
| Jacobse and Harskamp (2011) | Y | N | N | PY | N | Y | N | N | N | N | Y | N | N | Y | N | N | Critically low | High |
| Kaçar et al. (2021) | Y | N | N | N | N | Y | N | N | N | N | Y | N | N | Y | Y | Y | Critically low | High |
| Kalaian and Kasim (2017) | Y | N | N | PY | N | Y | N | N | N | N | Y | N | N | Y | N | N | Critically low | High |
| Karagöl and Esen (2018) | Y | N | N | PY | N | Y | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Karich et al. (2014) | Y | N | N | PY | N | Y | N | PY | Y | N | N | N | N | N | N | N | Critically low | High |
| Kim et al. (2021) | Y | N | N | PY | N | Y | N | Y | N | N | Y | N | N | Y | Y | N | Critically low | High |
| King-Sears et al. (2021) | Y | N | N | PY | Y | Y | N | PY | N | N | Y | Y | Y | Y | Y | Y | Critically low | High |
| Korpershoek et al. (2016) | Y | N | N | PY | N | Y | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Låg and Sæle (2019) | Y | N | N | PY | N | Y | N | N | Y | N | Y | N | Y | Y | Y | N | Critically low | Low |
| Leung (2015) | Y | N | N | PY | N | Y | N | N | N | N | Y | Y | Y | Y | Y | Y | Critically low | High |
| Linden et al. (2018) | Y | N | N | PY | Y | Y | PY | Y | Y | N | Y | Y | Y | Y | N | Y | Critically low | Low |
| May et al. (2021) | Y | Y | N | PY | Y | Y | N | Y | Y | N | Y | Y | Y | Y | Y | Y | Low | Low |
| Moore et al. (2018) | Y | Y | N | Y | Y | N | N | Y | Y | N | Y | Y | Y | Y | N | Y | Critically low | Low |
| Norris et al. (2019) | Y | Y | N | Y | Y | Y | N | Y | Y | N | Y | Y | Y | Y | Y | Y | Low | Low |
| Oh-Young and Filler (2015) | Y | N | N | PY | Y | N | N | PY | Y | N | Y | N | N | N | N | N | Critically low | High |
| Orhan (2019) | Y | N | N | PY | N | Y | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Petersen-Brown et al. (2019) | Y | N | N | PY | N | Y | N | PY | PY | N | Y | Y | Y | Y | Y | N | Critically low | Low |
| Phelps (2019) | Y | N | N | N | N | N | N | N | N | N | Y | N | N | Y | Y | Y | Critically low | High |
| Robbins et al. (2009) | Y | N | N | PY | Y | Y | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Saw and Han (2021) | Y | N | N | PY | N | N | N | Y | N | N | Y | N | N | Y | Y | Y | Critically low | High |
| Sayyah et al. (2017) | Y | N | N | PY | Y | N | N | Y | Y | N | Y | Y | Y | Y | Y | Y | Critically low | Low |
| Semerci and Batdi (2015) | Y | N | N | PY | N | Y | N | N | N | N | Y | N | N | Y | N | N | Critically low | High |
| Sneck et al. (2019) | Y | N | N | PY | Y | N | N | Y | Y | N | Y | Y | Y | Y | N | Y | Critically low | High |
| Steele et al. (2016) | Y | N | N | PY | Y | N | N | Y | Y | N | Y | Y | Y | Y | Y | N | Critically low | Low |
| Sugano and Nabua (2020) | Y | N | N | N | N | Y | N | N | N | N | Y | N | N | N | N | N | Critically low | High |
| Sung et al. (2016) | Y | N | N | PY | N | N | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Tan et al. (2017) | Y | Y | Y | PY | Y | Y | N | PY | Y | N | Y | Y | Y | Y | Y | Y | Low | Low |
| Tokac et al. (2019) | Y | N | N | PY | Y | N | N | PY | N | N | Y | N | N | Y | Y | Y | Critically low | High |
| Warren, 2012 | Y | N | N | Y | N | N | N | N | N | N | Y | N | N | Y | N | N | Critically low | High |
| Watson et al. (2017) | Y | Y | N | PY | Y | N | N | Y | Y | N | Y | N | N | Y | N | Y | Critically low | High |
| Wilson et al. (2019) | Y | N | N | PY | Y | N | N | Y | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Yang et al. (2020) | Y | N | N | Y | N | Y | N | N | N | N | Y | N | N | Y | Y | N | Critically low | High |
| Yorio and Ye (2012) | Y | N | N | N | N | N | N | PY | N | N | Y | N | Y | Y | N | N | Critically low | High |
| Zheng et al. (2016) | Y | N | N | N | Y | N | N | N | N | N | N | N | N | N | Y | N | Critically low | High |

†Seven critical domain. * Fisher's exact test. # Risk of bias: Lower (<2 critical flaws); Higher (>2 critical flaws).