London Review of Education

LONDON REVIEW OF EDUCATION

UCLPRESS

Special issue: ***Systematic reviews in education***

Research Article

Balancing workload with sensitivity to efficiently identify randomised controlled trials in an education systematic review

Claire Stansfield,[1,*] Alison O'Mara-Eves[1]

[1] EPPI Centre, UCL Social Research Institute, UCL Institute of Education, London, UK

* Correspondence: c.stansfield@ucl.ac.uk

Submission date: 22 December 2023; Acceptance date: 14 April 2025; Publication date: 21 May 2025

How to cite

Stansfield, C. and O'Mara-Eves, A. (2025) 'Balancing workload with sensitivity to efficiently identify randomised controlled trials in an education systematic review'. *London Review of Education*, 23 (1), 11. DOI: https://doi.org/10.14324/LRE.23.1.11.

Peer review

This article has been peer-reviewed through the journal's standard double-anonymous peer-review process, where both the reviewers and authors are anonymised during review.

Abstract

There is increasing interest in improving the efficiency of systematic review production, yet there is limited literature considering its application within the education field. This article analyses the study identification process adopted in a systematic review on effective teacher professional development, which identified 121 randomised controlled trials. It considers both human and technological inputs that aided production. It draws on project notes, an analysis of database sources and terminology used to identify randomised controlled trials, a retrospective evaluation of useful search terms and an analysis of using machine learning to reduce human workload during eligibility screening of citation records. Study identification was aided by four team processes (relating to ways of working and understanding the review context), the choice of information sources spanning education, psychology and economics research, and a variety of search terms for randomised controlled trials. The search resulted in 5,527 records identified from the main searches, and a further 3,614 records from forward and backward citation searching from the 121 included randomised controlled trials. Machine learning reduced screening workload, but implementation challenges included decisions on when to cease manual screening. In conclusion, carefully planned literature searches combined with machine learning to support eligibility screening can provide workload savings for sensitive study identification of randomised controlled trials in education. Improved reporting of randomised controlled trial design within research would aid these processes. Tools could also be developed to aid implementation of machine learning.

Keywords systematic review methods; literature searching; information retrieval; machine learning; text mining; citation screening; randomised controlled trials; citation searching

Introduction

Since the early 2000s, researchers have documented the challenges of efficiently and comprehensively identifying randomised controlled trials (RCTs) within education and social sciences for systematic reviews. Early work documented the challenges of retrieving relevant studies from bibliographic database searches and identifying research published outside peer-review journals (Petrosino et al., 2000; Turner et al., 2003). Such issues remain, and they have been compounded by the growth in primary research production over time, which makes study identification even more time-consuming, prompting the use of automation to support the review process (Gough and Thomas, 2016). The purpose of this study is to explore the study identification processes used in an education review – including both traditional (manual) and newer (semi-automated) approaches – to better understand the challenges and solutions in this discipline.

Background

The production of systematic reviews within the field of education is increasing. In January 2025, an international registry of protocols and published systematic reviews in education (IDESR, n.d.) contained over 490 entries for language education alone, of which a third are dated from 2020, despite containing reviews as far back as 1981. The Campbell Collaboration (n.d.) has published 50 systematic reviews since 2006, and the UK Educational Endowment Foundation (EEF, n.d.) toolkit of evidence contains over 50 evidence reviews since it started in 2011. These registries of education systematic reviews indicate an active research field. There is also demand for rapid reviews for informing policy decisions in education (Wollscheid and Tripney, 2021).

Despite a proliferation of education evidence syntheses, there is a dearth of recent literature focusing on the methods of study identification for systematic reviews within the education field. Specifically, there is a small evidence base for education reviews on which information sources to search and which search terms to use. This is a problem, because identifying the studies to analyse within a systematic review often requires a substantial amount of time. It calls for a literature search that is thorough enough to capture the research that exists, and for the search results to be screened against the eligibility criteria for the review.

Typically, the search strategy is tailored to limit the volume of research that should be manually screened for eligibility. Adopting machine learning to reduce the volume of manual screening may provide some flexibility in undertaking a literature search, and it could contribute to timely provision of systematic reviews. In a semi-automated prioritised screening approach, manual screening decisions train a machine classifier to rank the literature search results in order of likely relevance, and those that appear to be least relevant after repeated iterations of manual screening and machine classification are not screened. This reduces the manual screening workload. Machine learning for screening titles and abstracts is increasingly available within many systematic review software tools (Jimenez et al., 2022), and this approach has been used and evaluated particularly within the healthcare domain (O'Mara-Eves et al., 2015). However, its application to education reviews has not been evaluated. There is therefore a need to consider and share experiences of how study identification of education research can balance thoroughness with efficiency, and how it might be supported by machine learning.

Identifying studies, especially RCTs, in education

Education research is typically identified from a range of sources. To identify education research, MacDonald et al. (2024) advised searching beyond main subject resources (such as the education database ERIC), and searching databases in other research fields that are relevant to the research question (for example, psychology and others). In their context of undertaking multiple education reviews, Heck et al. (2024) found that searching discipline-focused international resources (either ERIC or Education Research Complete [EBSCO]) and national (German) databases was better than searching the multidisciplinary resource Web of Science. However, Pickup et al. (2018) found, for two systematic reviews of experimental studies on specific education topics, that searching multidisciplinary databases, checking references of key studies and reviews, and manual website browsing were important for identifying unique research in addition to subject-specific databases. Planning a literature search is challenging, and Wade et al. (2006) emphasised the importance of involving experienced information specialists in the literature searching stage.

Poor reporting quality is one potential barrier to study identification in general, but it appears to be magnified for identifying RCTs in education research. A survey of 89 RCTs, published in high-impact journal articles of education research, found that an indication of an RCT design was present in just 32 per cent of abstracts and 1 per cent of titles (Grant et al., 2013), although this may have since improved with availability of reporting standards, such as CONSORT-SPI (Grant et al., 2018). One problem may lie in conveying potentially complex methods using randomised controlled study designs within a brief abstract. For example, a majority of recent RCTs in education used a cluster-randomised design (Connolly et al., 2018), and the units of random assignment of an intervention may differ to the units being tested – for example, hierarchical designs where an intervention is targeted at a school, classroom or teacher, but outcomes are measured at the student level (Puma et al., 2009). Such complexity can be difficult to convey in the limited word count required by most journals. Poor reporting of research hinders discovery through browsing journals or searching databases, especially where it makes database indexing more difficult; this lack of discovery may ultimately contribute to research waste and duplication of research that has already been done (Layton and Clarke, 2016).

Regarding search terms for RCTs, the former Campbell Collaboration guidance (Kugley et al., 2017) advised taking a broad approach beyond searching for study types, and to experiment with different keywords 'such as *study*, *studies*, *evaluation*, *control group**, *random** etc.' (Kugley et al., 2017: Section 5.5.5; emphasis in original), while balancing inclusiveness with the time and budget available. For their scoping review of RCTs within the education field, Connolly et al. (2018) used a wide set of terms for title searching to be inclusive and parsimonious (that is, they used a broad range of terms, while trying to avoid multiple terms that did not retrieve anything new and useful).

Overall, there is no clear approach on the optimal information sources or search terms for identifying RCTs in education. Furthermore, we are not aware of any published work reflecting on the implementation of machine learning for eligibility screening in this field. This study contributes to this area by exploring the study identification processes used in an education review to better understand the challenges and solutions in this discipline. It considers both human and technological inputs that aided production. As part of doing so it aims to understand:

1. elements that helped with developing the search strategy to identify RCTs on teacher professional development
2. which database sources were most useful for finding relevant RCT studies
3. what search terms help identification of RCTs
4. the benefits and challenges of using semi-automated screening on the database search results
5. the benefits and challenges of two semi-automated screening approaches on the results from citation searching.

Methods

Design

This study describes the authors' experiences of study identification from undertaking one systematic review (Sims et al., 2021, 2023), and it is strengthened by retrospective analyses of the methods used. It also draws on a sample of 947 records included in a scoping review of RCTs in education (Connolly et al., 2018), which was analysed after the Sims et al. (2021, 2023) systematic review was completed.

Data

The original systematic review analysed the characteristics of effective teacher professional development (Sims et al., 2021, 2023). The review was undertaken in nine months, and it identified 121 relevant RCTs published in English between 2005 and 2020, of which 104 contained sufficient data on effect size for undertaking meta-analysis. Studies were identified by searches of 10 scholarly databases of research and other sources to yield 5,527 records for screening, after de-duplication. Machine learning (based on text-mining) was used to iteratively rank the references in order of likely relevance so that 57 per cent (3,140) of these records were screened manually by the review team, with the remainder unscreened. EPPI-Reviewer software (Thomas et al., 2022) was used to store, screen and analyse the research citations identified from the search. Manual searches of relevant websites were undertaken by reviewers with prior topic knowledge of the review focus. In addition, a rapid approach to forward and backward citation searching was undertaken on the studies in the review synthesis that were available in Microsoft Academic (n = 108). This involved searches of Microsoft Academic within the EPPI-Reviewer interface to yield 3,614 new records after removing duplicates. The search results were ranked by relevance using two separate machine-learning classifiers, and 11 per cent of the 3,614 records ranked as most relevant were manually screened. The review team comprised education research and practice experts, an experienced systematic reviewer (AOE) and an experienced systematic review information specialist (CS) (the latter two being authors of this study). The methods are reported in further detail in the original review report (Sims et al., 2021).

The data for the present study largely come from the bibliographic records stored in EPPI-Reviewer that were collated as part of the original review (Sims et al., 2021, 2023). This database was used to run simulations of prioritised screening (that is, the ranking of records in order of most likely to least likely to be relevant to the review using machine-learning techniques), which generated new datasets (described below). Finally, data for the reflective work came from notes and emails recorded by the research team during the original review.

Analysis

Five approaches were taken to address the research aims:

1. *Reflection on search strategy development to identify RCTs on teacher professional development:* Reflection was supported by notes saved during the search strategy development phase and rereading of emails between the review team and the information specialist. This primarily addressed the first aim: to identify elements that helped with developing the search strategy.
2. *Analysis of sources from which the RCTs were retrieved:* Database sources of 121 citation records of RCTs included within the systematic review (Sims et al., 2021, 2023) were identified from the duplicate-report function within EPPI-Reviewer and analysed in Excel. The data were checked for duplicates. The citation records from the original searches undertaken during November 2020 were grouped into those identified from the primary database used, ERIC (EBSCOhost), and those only identified elsewhere. The records that were not found through the original searches of ERIC were checked to see if they were present within ERIC by searching on phrases within their titles. This addressed the second research aim: to identify database sources that were most useful for finding relevant RCT studies.
3. *Retrospective evaluation of useful search terms for identifying randomised controlled trials:*
    a. *Controlled vocabulary used by electronic databases:* Bibliographic records of the RCTs found from ERIC were checked for ERIC descriptors that could indicate an RCT study design. For the

RCTs found in other databases (not present in ERIC), their records within the databases were checked for any controlled terms associated with an RCT design. Analysis was supported by using a frequency of subject headings function in EndNote reference management software (EndNote Team, 2013)

b. *Free text terms:* Terms in the titles and abstracts relating to an RCT study design were extracted from two datasets: 121 records and 947 records used for a scoping review of RCTs in education (Connolly et al., 2018) (obtained after removing records without abstracts, duplicates and spell-checking, and split into citations using TextWedge). Text analysis was undertaken iteratively using Antconc (Anthony, 2023) to determine frequency of terms per record, searching the records within EndNote with identified terms, and scanning remaining records without these terms for suitable search terms. This addressed the third aim: determining which search terms help identification of RCTs.

4. *Benefits and challenges of using semi-automated screening on the database search results:*
   a. The process of semi-automated prioritised screening is described in the findings, along with a reflection on implementing this process during the review
   b. Retrospective simulations were undertaken of a prioritised approach to screening title and abstracts, and both before and after records from backward and forward citation searching were incorporated (including the citation search results reduced the yield of included studies from 6 per cent to 4 per cent at title-abstract screening stage). A tool within EPPI-Reviewer simulated the prioritised screening process involving continuous active learning. The tool randomly samples 'seed' records with two include and two exclude decisions from the original review and runs a standard logistic regression algorithm over every 25 records 'screened'. This is undertaken 10 times to account for variation in the initial 'seed' records. The output data were imported into Excel to calculate average screening volume for recall of the included study records at 90, 95, 99 and 100 per cent recall.
5. *Benefits and challenges of two semi-automated screening approaches of the results from citation searching:* Two approaches were used to rank the results of forward and backward citation searching by likely relevance, and the most relevant were manually screened. One approach was based on the screening decisions in the review, and the second on relevance-ranking of RCT-design (which is explained further in the findings and discussion section). We compared both approaches from retrospectively analysing the relevance ranking of the records.

Findings and discussion

Elements that helped with developing the search strategy

Four main elements helped the information specialist (CS) to develop the search strategy:

1. *Drawing on previous work:* Terms for both professional development and RCTs needed to be developed for the database search. Enquiries were made to the Campbell Information Retrieval Methods Group to identify approaches for finding RCTs. The strategy to search for RCTs was developed using terms that had previously been applied to a scoping review of RCTs across education research (Connolly et al., 2018). These terms were modified to avoid obtaining unnecessarily high volumes of search results (which was identified during test searches), and additional terms were introduced to reflect the context of the review (for example, using terms for 'randomised' near to the terms for 'teachers' or 'educators'). The search terms were also informed by investigating relevant references from three published reviews on teacher professional development (Filges et al., 2019; Kennedy, 2016; Kraft et al., 2018) and from previous experience of evaluating search terms to identify controlled trials for a public health register of research (TROPHI, EPPI Centre).
2. *Knowing in advance that not all records would be manually screened:* The literature search aimed to be comprehensive, and bibliographic database searches aimed to achieve a pragmatic balance of being focused on the topic with sensitivity to capture relevant records. Although it was planned to apply semi-automated prioritised screening and not manually screen the least relevant search results, at the literature searching stage, it was neither clear how many would need to be screened

manually, nor if there would be a clear cut-off threshold for ceasing manual screening. Therefore, while using this method allowed some flexibility on the volume of records obtained from the database search, there were still boundaries on the acceptable volume of search results. Searching the ERIC database alone yielded 3,808 records, and the ERIC database search was used as a template for searching the other databases.

3. *Team discussions:* The information specialist engaged with the rest of the team on the approach for screening the title and abstract records. The topic experts within the review team initially favoured a narrow approach for identifying RCTs, with the expectation that authors of RCTs would want to clearly state their method in the title and abstract. However, exploratory searches in ERIC found records showed that title and abstract could not be relied upon to indicate an RCT design. These searches also showed that relying on controlled vocabulary for identifying RCTs was not sufficient. It was important for the information specialist to collaborate with the whole review team to gain understanding of the topic and expectations of volume, to illustrate challenges identified through the exploratory searches and to discuss how the results would be screened.
4. *Iteratively designing the search:* The search was developed through iterative testing, and through sharing search terms and findings with the team (which is conventional practice).

Which database sources were most useful for finding relevant RCT studies?

The number of total and unique relevant records identified in each database source are shown in Table 1. Out of 121 records, 83 (68.6 per cent) were found from ERIC (EBSCOhost). A further 7 were present in this platform of ERIC, although missed by the searches owing to either not containing the search terms used for professional development (in the case of 5 records), or being added to the database after the original search was conducted (2 records). However, the results appear variable, for at least 2 other records were on ERIC (EBSCOhost) at one stage but could not be located on a later date, even though they were present on the public ERIC platform at eric.ed.gov (that is, it seems that the EBSCOhost platform did not consistently hold records that were on the public version of the ERIC database, based on the searches to check the presence of individual records).

Of the 38 (31.4 per cent) records not identified from ERIC searches, 17 were identified from manual searching and browsing of website sources, and 21 were identified from the following databases: PsycInfo (OVID) (*n* = 12), Education Abstracts (H.W. Wilson, EBSCOhost) (*n* = 8), Econlit (EBSCOhost) (*n* = 2), British Education Index (EBSCOhost) (*n* = 2), Educational Administration Abstracts (EBSCOhost) (*n* = 2), Microsoft Academic Graph (MAG) (EPPI-Reviewer interface) (*n* = 2) and Australian Education Index (ProQuest) (*n* = 1). Of these 21 records, 6 were in multiple databases.

MAG was only used for forward and backward citation searching. Forward citation searching was important for retrieving 2 records that were missed from the database searches as they did not have search terms or indexing that identified them as RCTs (these records were present within PsycInfo but not clearly retrievable as an RCT). No additional records were identified from backward citation searching in MAG.

Table 1. Records included in Sims et al. (2021, 2023), identified by database searches

Source	Relevant records	Unique (*n* = 15)	Controlled terms for RCTs (present within the identified study records only)	Details of uniquely identified items and publication source
ERIC (EBSCO)	83 (out of 90 known to exist in database)	(not determined)	Control groups Experimental groups Outcome measures Program evaluation RCTs	(not determined)

Table 1. ***Cont.***

Source	Relevant records	Unique (*n* = 15)	Controlled terms for RCTs (present within the identified study records only)	Details of uniquely identified items and publication source
Study records not identified by the original systematic review searches in ERIC				
PsycInfo (OVID)	12	7	RCTs Methodology: Clinical trial	Correnti et al. (2021) (*Reading Research* Quarterly) Connor et al. (2007) (Science) Haring (2013) (doctoral thesis) Landry et al. (2017) (*Early Childhood Research* Quarterly) Mazzie (2008) (doctoral thesis) Reinke et al. (2018) (*Prevention* Science) Snow et al. (2014) (*Advances in Speech Language Pathology*)
Education Abstracts (H.W. Wilson, (EBSCOHost)	8	2	None	Ansari and Pianta (2018) (*Early Childhood Research* Quarterly) Arteaga et al. (2019) (*Evaluation Review*)
Econlit (EBSCOHost)	2	2	None	Jacob (2017) (*Labour* Economics) Jerrim and Vignoles (2016) (*Economics of Education Review*)
BEI (EBSCOHost)	2	1	None	Anonymous (2018) (*Education Journal* – report of trial by EEF)
Australian Education index (ProQuest)	1	1	RCTs Program evaluation	Dix et al. (2018) (report)
MAG (citation searching, EPPI-Reviewer)	2	2	None	Castro et al. (2017) (*Early Childhood Research Quarterly*) Prast et al. (2018) (*Learning and Instruction*)
Educational Administration Abstracts (EBSCOHost)	2	0	None	

What terms support identification of RCTs?

Free-text searches of title, abstract and author keywords are essential for searching on RCT design in the bibliographic databases, as only around one-third of the records would be identified using the controlled vocabulary detailed in Table 1, Column 4. Useful controlled terms that indicate an RCT design were limited to records from ERIC, PsycInfo and the Australian Education Index.

Text analysis of titles and abstracts from the 121 RCTs (from Sims et al., 2021, 2023) show the importance of being expansive in using terms related to RCT design and going beyond synonyms for 'randomised'. Analysis of the second dataset of 947 RCTs across different education subfields (from Connolly et al., 2018) strengthens this conclusion, and both datasets were used to develop a set of terms and phrases to test in future searches (available from Stansfield, 2025). Five key findings on these search terms are described here:

1. Use terms describing the control arms or intervention groups without an indication of 'random', for example: 'treatment/experimental groups', 'experimental conditions', 'treatment as usual' and 'business as usual'. This includes using alternative terms for controls such as 'control' and 'comparison' (one abstract used the term 'contrast group'). However, the word 'control' (as in 'control students' or 'control teachers') also yields records pertaining to behaviour control.
2. Use different terms for potential types of grouping (for example, 'student', 'classroom', 'teachers', 'school', 'condition').
3. Use phrases such as 'randomly allocated', 'random assignment' or 'randomly divided'.
4. Use plurals for trial and related words to identify records that describe multiple studies.
5. Consider searching on the abbreviation 'cRCT' for a 'cluster randomised controlled trial', although this may also locate other uses of this abbreviation, such as criterion-referenced competency tests.

The benefits and challenges of using semi-automated screening on the database search results

Machine learning using text-mining within the EPPI-Reviewer review management software was used to rank all records in order of potential relevance, known as prioritised screening. This semi-automated approach involves active learning, whereby the accuracy of relevance ranking is improved by the manual screening decisions of reviewers, so that the machine-learning system 're-ranks' at differing screening time-points (EPPI Centre, n.d.). To complete the review within the nine-month deadline, using prioritised screening and stopping screening at an appropriate stage was important. Manual screening of title and abstract records was undertaken on 3,140 records, including those that only had titles and all those identified from manual web searching and browsing. The unscreened records remaining when screening ceased was 2,386. Screening results from citation searching is considered separately in the next section.

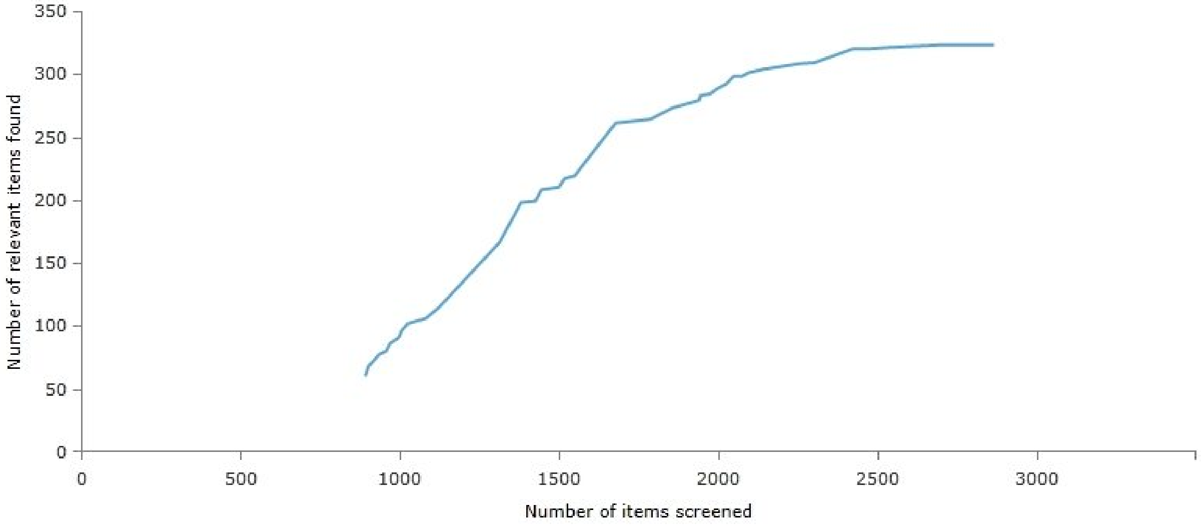
Application of prioritised screening can be beneficial when developing search strategies, although it does not remove challenges in designing a suitable search. A benefit is that search results are not as closely bound to the level of resources available to screen, compared with a traditional search. However, a challenge is that it is unknown at this stage how many records from the search results will need to be manually screened. This uncertainty requires the searches to avoid relatively excessive volumes of results where possible. Retrospective prioritised screening simulations on different dates show that increasing the volume of irrelevant records by 60 per cent would have increased the volume of records to be screened by at least 49 per cent (1,011 records) and 65 per cent (1,125 records) for 95 per cent recall. While these figures are only a tentative indication of impact, they show how variation at the search stage may increase the number needed to be manually screened.

Another benefit of this approach was reduction in time spent screening, as 43 per cent (2,386) of the database search results were not screened. Manual screening of the records required careful reading owing to the diverse terminology of the topic, so screening relatively modest volumes of records was time-consuming. For example, of the 347 records included at abstract screening stage, 79 were flagged for a second opinion for inclusion, indicating that inclusion was not clear. All of these were incorporated within the full-text retrieval stage. Therefore, we consider this approach as timesaving, although note that time spent screening was not measured in the present study. Further, screening low-relevance records may require less time than screening higher relevance records (Stansfield et al., 2022). Also, a decision was taken to manually screen all the title-only records (that is, those without abstracts) from the database search results because it was uncertain how well screening prioritisation would work for them. While this decision may sometimes have implications for reducing workload savings, it was only around 120 records in this case.

The decision to cease manual screening is a challenge in the absence of commonly accepted stopping criteria. In this case, stopping criteria were based on two factors: observing a plateau in the screening inclusion rate (Figure 1) and exceeding a predicted number of includes. This prediction was achieved by double-screening a random sample of 15 per cent (*n* = 863) of the records before applying machine learning, which was determined to be a sufficiently powered random sample (the methods are detailed in Appendix 1 of Sims et al., 2021). At this stage, 344 records were seen as potentially relevant, which corresponds to an inclusion rate of 11 per cent for the records manually screened, and 6 per cent of the total records available to screen at that stage. These values informed the decision to stop manual screening, but this approach is a crude estimate and has not been formally evaluated; the approach was taken in the absence of any consensus or guidance around appropriate stopping criteria at the time of the study.

A decision not to conduct full double-screening is both a benefit (in terms of work saved) and a challenge (in terms of uncertainty around manual screening errors). Within the team, there were three main screeners and a fourth who helped with title-only screening, resolving disagreements and screening of the records from citation searching. They were all familiar with education research, and two were experts in the review topic area. This supported an approach that involved single-screening after ensuring that there was sufficient consistency from double-screening. Although 15 per cent (*n* = 863) of the available records were screened by two reviewers, the manual process is not without error. A recent estimate from screening titles and abstracts in health literature estimated that screening errors led to 13 per cent of missed studies for single screeners, and 3 per cent in dual screeners (Gartlehner et al., 2020). Belur et al. (2021) analysed inter-rater reliability in screening for two reviews of crime-related research: they observed variation in subjective decision-making both by individuals at different time-points of screening and between screeners, and individuals differed in their responses to ambiguous abstracts. It is not known how prioritised screening can help or hinder the reliability of screening decisions, and reliability might vary through screening time-points. For example, in the early stages of screening, the screener is becoming familiar with the variety of abstract records being presented to them that are ranked as high relevance. As screening progresses, the relevant records are less frequent, and records potentially may be more ambiguous or differ from the (now-familiar) earlier records in other ways. It is hoped that future improvements in machine-learning tools for screening could provide support to aid quality assurance of single-screening processes.

Figure 1. Rate of identification of relevant items through prioritised screening



Benefits and challenges of semi-automating forward and backward citation searching

MAG within EPPI-Reviewer was used as a rapid approach to undertake forward and backward citation searching of the originally included studies in the evidence synthesis. This approach required matching the included studies with their records in MAG (of which 108 were able to be matched), applying the functions 'cited by' (for forward citation searching) and 'bibliography' (for backward citation searching), and importing the records to the review management database. The MAG database has since become obsolete and has been replaced by OpenAlex. We estimate the time to match, search and import the search results was around 65 minutes (based on replicating these processes in October 2023).

The original citation-searching searches yielded 3,614 records (2,164 from forward citations and 1,450 from backward citations, respectively). This is comparable to the 3,140 manually screened for the entire review up to the point of these searches. Given that one benefit of citation searching is to mitigate deficiencies from text-based searches in databases, it was unclear whether relying on prioritised screening based on the screening decisions of the database searches would be sufficient. Such an approach might introduce a 'more-of-the-same bias', known as 'hasty generalisation' (O'Mara-Eves et al., 2015), whereby relevant records that use different terminology to those in included records would be automatically de-prioritised. Therefore, two different machine-learning classifiers were applied on this corpus of records. First, a classifier was trained on all the included studies and the excluded studies that were excluded on intervention type, at the title and abstract screening stage. This criterion was broader than the exclusions used for training prioritised screening as it did not train on exclusions for population or study design. Second, the Cochrane RCT classifier within EPPI-Reviewer (which was trained on over 280,000 records of health RCTs; Marshall et al., 2018) was used to look for terms related to RCTs. For both classifiers, a rule was applied whereby any record that was given a relevance score of $\geq$50 per cent was manually screened on title and abstract. This threshold was an arbitrary value determined for pragmatic purposes and, by doing so, 11 per cent (*n* = 412 out of 3,614) were manually screened to yield 12 records included at title and abstract, of which two were included at full text. The two included were identified from the first approach, using a classifier trained on included studies, and screening 359 records. Applying the second RCT classifier resulted in screening a further 53 records, which yielded 1 record considered potentially relevant that was subsequently excluded on full text (this record was just below the threshold for manual screening with the first classifier). The second classifier was used as an additional pragmatic and exploratory process, to try to rapidly mitigate 'more-of-the-same bias'. The method of using a study-design classifier did not prove fruitful in this case, apart from enabling an additional set of records to be selected for screening. It tentatively illustrates the need for topic-differentiation, if it is to be used for efficient screening gains.

However, it should be noted that conventional use of this classifier would have required a much lower screening threshold for their discovery (EPPI Centre, n.d.) and would have involved further manual screening of 1,460 records, which were lower ranked on topic, according to the first classifier. Within EPPI-Reviewer, two RCT classifiers are available, 'Cochrane RCT', and 'original RCT', which differ slightly (owing to their machine classification methods). The 'original RCT' is considered more suitable for use than the Cochrane RCT classifier where a user-determined cut-off threshold is applied (EPPI Centre, n.d.), although it appears less sensitive in this case and would only have identified the one record that was identified by the Cochrane RCT classifier at a lower threshold. Both these classifiers are trained on the same health data, and their recall of RCTs in education overall is lower. We estimate, based on the 121 included records in the synthesis, that the Cochrane RCT classifier (which is calibrated to 99 per cent recall) would have found 97 per cent (*n* = 117), while the original RCT classifier (using a threshold score of above 10) would have found 87 per cent (*n* = 105). This indicates the need for topic-related training data and further evaluations within the education domain.

Overall, the approach using MAG (now incorporated into OpenAlex) was successful. It identified two additional studies and saved time for retrieving and managing results from forward and backward citations for 121 records, as the process is largely automated. However, it relies on both the record being present in MAG/OpenAlex and for sufficient data to exist within the citation record. In this case, the approach retrieved a high volume of records, comparable with those collected for the database searches.

Citation searching within a traditional study identification process helps verify the utility of the database search strategy. In this case, given that only two further studies were identified by this method, it supports our confidence in the utility of the database searches. Studies have shown that citation searching results vary across citation databases (for example, Gusenbauer, 2024). The recent TARCiS statement (Hirt et al., 2024) introduces recommendations on the conduct and reporting of citation searching. It encourages consideration of using two citation indexes to achieve greater coverage of available literature. However, it also acknowledges that citation searching on a large number of records could yield too many records to screen, and, in such cases, a selected sample of references could be used as the starting point for citation searching (provided there is justification of the sampling method). In the Sims et al. (2021, 2023) systematic review, citation searching was undertaken on all available references in one resource, and sampling via machine learning was applied to the results of citation searching. Although this method requires further verification, it could be a promising alternative approach.

Conclusions and recommendations

Issues to consider when identifying studies for RCTs in education

This study revealed various features which may be useful in enhancing study identification in education systematic reviews. These are:

- Team processes:
    - Knowing in advance that not all the records would be manually screened, so that the search strategy is not entirely dependent on volume of results that can be manually screened.
    - Drawing on previous work to identify more extensive terminology in an efficient way.
    - Engaging with the team members involved in screening regarding how they were planning to screen the records, particularly in terms of the explicitness of terminology for key screening criteria in titles and abstracts (for example, the terminology of RCTs if records are being screened based on description of RCT study design).
    - Taking an iterative approach to develop the search strategy (which is conventional practice).
- Technical approaches:
    - Using ERIC and manual searching and browsing of websites and repositories of education research to identify the bulk of unique records, complemented by searches in other databases. For this topic, PsycInfo was important to find both journal articles and doctoral theses not found from the other searches (*n* = 7) and using the Australian Education Index, British Education index, Econlit, Education Abstracts and citation searching increased comprehensiveness. Unfortunately, at the time of publishing this article, we are aware of planned changes to new content added to ERIC that will involve a significant reduction in the journals that are indexed. Therefore, alternative resources will need to be searched to counter this loss.
    - Using a wide range of terms to identify RCTs, including terms relating to random, grouping, and clustering, and, if possible, using broader terms to capture studies with comparison designs.
    - Careful screening is needed to judge that a study potentially uses an RCT design.

Using technology for more efficient, higher quality systematic reviews

This study highlights various benefits and challenges of implementing machine learning using text-mining processes for screening and semi-automating citation searching:

- Semi-automated prioritised screening appeared to reduce time spent screening because only 56.8 per cent of the records (those that were most likely to be relevant) were manually screened.
- At the start, it is unknown how many records will need to be manually screened, so searches should still aim for a degree of precision. Conversely, the search results may be less bound to the level of resources available to screen if there is an expectation that manual screening can stop early. We expect that screening takes the bulk of the time spent on study identification, and if there is a time–resource trade-off, then it seems sensible to reduce screening rather than curtail searches.
- Searching more databases and browsing websites may require additional time and thinking involved in navigating each source, although it may not necessarily increase the screening workload, and it is therefore recommended if possible.

- However, citation searching on a large set of relevant references did increase the workload. Citation searching using semi-automation appears to be an efficient approach to identify further studies, but more investigation and guidance is needed on thresholds for using classifiers to narrow the pool for manual screening.
- Given the time-limited nature of the systematic review and increasing volume of research, using technology to reduce workload is appealing, although a challenge is to maintain rigour of implementation as methods of application are being developed.
- The decision to cease manual screening is a challenge because there are no clear standards or guidance on stopping rules (Callaghan et al., 2024).
- As with all systematic reviews, there is autonomy by review teams to develop and implement search strategies and screening protocols. The use of technology requires thought and guidance on implementation.
- Prioritised screening may interact with human screener decisions by potentially predisposing the reviewer to a particular include or exclude decision. This may have implications for single screening, and it would benefit from further evaluation.
- We are not aware of an RCT machine classifier for education research. Such a classifier would need to use sufficient training data from the education field to achieve high recall.

Final remarks

This study informs the methods of study identification for RCTs in education and reflects on the current and potential benefits of machine learning to improve efficiency. Although the work focuses on one case study, the quantity of RCTs and the use of additional analyses strengthens its findings for potential use in other reviews, and it adds to the paucity of literature that exists. However, more studies are needed within the education domain to strengthen methods of study identification.

This study uses the ERIC database as the primary database source, and the majority of literature for the systematic review on which it focuses was retrievable from ERIC. However, the journal content is expected to reduce significantly during April 2025. ERIC (n.d.), which is an acronym of Education Resources Information Center, was established in the 1960s by the US Federal Department of Education following a recommendation for an information service that encompasses all education research to inform theory and practice. Until April 2025 it contained over 2 million records and was available worldwide, both in open access form and through subscription database platforms. This context justifies our approach to this study; however, the reduction in content illustrates that there is a need to examine the overlap of other information resources in relation to the culled journals.

The scope for both human and machine error during screening, and determining suitable thresholds for ceasing manual screening should be carefully balanced with the need to provide timely and rigorous evidence to inform decision-making. We recommend improved tools and guidance to support the appropriate use of prioritised screening and stopping thresholds, including further investigation into methods of combining citation searching and prioritised screening. We add to the call for reporting and indexing to label RCT research by authors and database providers in education, so that this can aid study identification and improve machine learning and prioritisation.

Acknowledgements

Our thanks go to Samuel Sims, the principal investigator of the systematic review on teacher professional development that is referred to in this article, and to Harry Fletcher-Wood and Sarah Cottingham, who also contributed to the study identification process for this review. We also thank Ciara Keenan for sharing RCT records in education.

Declarations and conflicts of interest

Research ethics statement

Not applicable to this article.

Consent for publication statement

Not applicable to this article.

Conflicts of interest statement

The authors declare no conflicts of interest with this work. All efforts to sufficiently anonymise the authors during peer review of this article have been made. The authors declare no further conflicts with this article.

References

Anonymous. (2018) 'Trial of embedding formative assessment by the education endowment foundation'. *Education Journal*, (348), 16.

Ansari, A. and Pianta, R.C. (2018) 'Effects of an early childhood educator coaching intervention on preschoolers: The role of classroom age composition'. *Early Childhood Research Quarterly*, *44*, 101–13. [CrossRef]

Anthony, L. (2023) *AntConc (Version 4.2.4)* [Computer software]. Waseda University. Accessed 15 April 2025. https://www.laurenceanthony.net/software.

Arteaga, I., Thornburg, K., Darolia, R., and Hawks, J. (2019) 'Improving teacher practices with children under five: Experimental evidence from the Mississippi Buildings Blocks'. *Evaluation Review*, *43* (1/2), 41–76. [CrossRef]

Belur, J., Tompson, L., Thornton, A. and Simon, M. (2021) 'Interrater reliability in systematic review methodology: Exploring variation in coder decision-making'. *Sociological Methods and Research*, *50* (2), 837–65. [CrossRef]

Callaghan, M., Müller-Hansen, F., Bond, M., Hamel, C., Devane, D., Kusa, W., O'Mara-Eves, A., Spijker, R., Stevenson, M., Stansfield, C., Thomas, J. and Minx, J.C. (2024) 'Computer-assisted screening in systematic evidence synthesis requires robust and well-evaluated stopping criteria'. *Systematic Reviews*, *13*, 284. [CrossRef]

Campbell Collaboration. (n.d.) *Research Evidence*. Accessed 23 January 2025. https://www.campbellcollaboration.org.

Castro, D.C., Gillanders, X.C., Franco, M., Bryant, D.M., Zepeda, T., Willoughby, M. and Méndez, L.I. (2017) 'Early education of dual language learners: An efficacy study of the Nuestros Niños School Readiness professional development program'. *Early Childhood Research Quarterly*, *40*, 188–203. [CrossRef]

Connolly, C., Keenan, C. and Urbanska, K. (2018) 'The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016'. *Educational Research*, *60* (3), 276–91. [CrossRef]

Connor, C.M., Morrison Frederick, J., Fishman, B.J., Schatschneider, C. and Underwood, P. (2007) 'Algorithm-guided individualized reading instruction'. *Science*, *315* (5811), 464–65. [CrossRef]

Correnti, R., Matsumura, L.C., Walsh, M., Zook-Howell, D., Bickel, D. and Yu, B. (2021) 'Effects of online content-focused coaching on discussion quality and reading achievement: Building theory for how coaching develops teachers' adaptive expertise'. *Reading Research Quarterly*, *56* (3), 519–58. [CrossRef]

Dix, K., Hollingsworth, H. and Carslake, T. (2018) *Thinking Maths: Learning Impact Fund Evaluation Report: Evaluation report and executive summary*. Sydney: Social Ventures Australia (SVA).

EEF. (n.d.) 'Evidence reviews'. Accessed 12 December 2024. https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews.

EndNote Team. (2013). *EndNote 21* [Computer software]. Philadelphia: Clarivate.

EPPI Centre. (n.d.) 'Machine learning functionality in EPPI-reviewer'. Accessed 31 October 2023. https://eppi.ioe.ac.uk/CMS/Portals/35/machine_learning_in_eppi-reviewer_v_7_web_version.pdf.

ERIC (Education Resources Information Center). (n.d.) *ERIC Through the Decades*. Accessed 2 April 2025. https://eric.ed.gov/pdf/ERIC_Through_the_Decades.pdf.

Filges, T., Torgerson, C., Gascoine, L. and Dietrichson, J. (2019) 'Effectiveness of continuing professional development training of welfare professionals on outcomes for children and young people: A systematic review'. *Campbell Systematic Reviews*, *15* (4), e1060. [CrossRef]

Gartlehner, G., Affengruber, L., Titscher, V., Noel-Storr, A., Dooley, G., Ballarini, N. and König, F. (2020) 'Single-reviewer abstract screening missed 13 percent of relevant studies: A crowd-based, randomized controlled trial'. *Journal of Clinical Epidemiology*, *121*, 20–8. [CrossRef]

Gough, D. and Thomas, J. (2016) 'Systematic reviews of research in education: Aims, myths and multiple methods'. *Review of Education*, *4* (1), 84–102. [CrossRef]

Grant, S.P., Mayo-Wilson, E., Melendez-Torres, G.J. and Montgomery, P. (2013) 'Reporting quality of social and psychological intervention trials: A systematic review of reporting guidelines and trial publications'. *PLoS ONE*, *8* (5), e65442. [CrossRef]

Grant, S., Mayo-Wilson, E., Montgomery, P., MacDonald, G., Michie, S., Hopewell, S. and Moher, D. (2018) 'CONSORT-SPI 2018 explanation and elaboration: Guidance for reporting social and psychological intervention trials'. *Trials*, *19*, 406. [CrossRef]

Gusenbauer, M. (2024) 'Beyond Google Scholar, Scopus, and Web of Science: An evaluation of the backward and forward citation coverage of 59 databases' citation indices'. *Research Synthesis Methods*, *15* (5), 802–17. [CrossRef]

Haring, C.D. (2013) 'The effects of coaching on teacher knowledge, teacher practice and reading achievement of at-risk first grade students'. PhD thesis, University of Texas at Austin, Austin, TX, USA.

Heck, T., Keller, C. and Rittberger, M. (2024) 'Coverage and similarity of bibliographic databases to find most relevant literature for systematic reviews in education'. *International Journal on Digital Libraries*, *25* (2), 365–76. [CrossRef]

Hirt, J., Nordhausen, T., Fuerst, T., Ewald, H. and Appenzeller-Herzog, C. (2024) 'Guidance on terminology, application, and reporting of citation searching: The TARCiS statement'. *BMJ*, *385*, e078384. [CrossRef]

IDESR (International Database of Education Systematic Reviews). (n.d.). 'Welcome to the International Database of Education Systematic Reviews'. https://idesr.org/ Accessed 2 April 2025.

Jacob, B. (2017) 'When evidence is not enough: Findings from a randomized evaluation of Evidence-Based Literacy Instruction (EBLI)'. *Labour Economics*, *45*, 5–16. [CrossRef]

Jerrim, J. and Vignoles, A. (2016) 'The link between East Asian "mastery" teaching methods and English children's mathematics skills'. *Economics of Education Review*, *50*, 29–44. [CrossRef]

Jimenez, R., Lee, T., Rosillo, N., Cordova, R., Cree, I.A., Gonzalez, A. and Ruiz, B.I.I. (2022) 'Machine learning computational tools to assist the performance of systematic reviews: A mapping review'. *BMC Medical Research Methodology*, *22*, 322. [CrossRef]

Kennedy, M.M. (2016) 'How does professional development improve teaching?'. *Review of Educational Research*, *86* (4), 945–80. [CrossRef]

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

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A.M.K., Hammerstrøm, K. and Sathe, N. (2017) 'Searching for studies: A guide to information retrieval for Campbell systematic reviews'. *Campbell Methods Guides*, *13* (1), 1–73. [CrossRef]

Landry, S.H., Zucker, T.A., Williams, J.M., Merz, E.C., Guttentag, C.L. and Taylor, H.B. (2017) 'Improving school readiness of high-risk preschoolers: Combining high quality instructional strategies with responsive training for teachers and parents'. *Early Childhood Research Quarterly*, *40*, 38–51. [CrossRef]

Layton, D.M. and Clarke, M. (2016) 'Lost in translation: Review of identification bias, translation bias and research waste in dentistry'. *Dental Materials*, *32* (1), 26–33. [CrossRef]

MacDonald, H., Comer, C., Foster, M., Labelle, P.R., Marsalis, S., Nyhan, K., Premji, Z., Rogers, M., Splenda, R., Stansfield, C. and Young, S. (2024) 'Searching for studies: A guide to information retrieval for Campbell systematic reviews'. *Campbell Systematic Reviews*, *20* (3), e1433. [CrossRef]

Marshall, I., Storr, A.N., Kuiper, J., Thomas, J. and Wallace, B.C. (2018) 'Machine learning for identifying randomized controlled trials: An evaluation and practitioner's guide'. *Research Synthesis Methods*, *9* (4), 602–14. [CrossRef]

Mazzie, D.D. (2008) 'The effects of professional development related to classroom assessment on student achievement in science'. PhD thesis, University of South Carolina, Columbia, SC, USA.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. and Ananiadou, S. (2015) 'Using text mining for study identification in systematic reviews: A systematic review of current approaches'. *Systematic Reviews*, *4* (1), 5. [CrossRef]

Petrosino, A.R.F., Boruch, C., Rounding, S., McDonald, S. and Chalmers, I. (2000) 'The Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR) to facilitate the preparation and maintenance of systematic reviews of social and educational interventions'. *Evaluation and Research in Education*, *14* (3/4), 206–19. [CrossRef]

Pickup, D.I., Bernard, R.M., Borokhovski, E., Wade, A.C. and Tamim, R.M. (2018) 'Systematically searching empirical literature in the social sciences: Results from two meta-analyses within the domain of education'. *Russian Psychological Journal*, *15* (4), 245–65. [CrossRef]

Prast, E.J., Van de Weijer-Bergsma, E., Kroesbergen, E.H. and Van Luit, J. (2018) 'Differentiated instruction in primary mathematics: Effects of teacher professional development on student achievement'. *Learning and Instruction*, *54*, 22–34. [CrossRef]

Puma, M.J., Olsen, R.B., Bell, S.H. and Price, C. (2009) 'Technical methods report: What to do when data are missing in group randomized controlled trials – 2. Randomized controlled trials (RCTs) in education and the problem of missing data'. NCEE 20090049. Accessed 15 September 2023. https://ies.ed.gov/ncee/pubs/20090049/section_2.asp.

Reinke, W.M., Herman, K.C. and Dong, N. (2018) 'The Incredible Years Teacher Classroom Management Program: Outcomes from a group randomized trial'. *Prevention Science*, *19* (8), 1043–54. [CrossRef]

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., Van Herwegen, J. and Anders, J. (2023) 'Effective teacher professional development: New theory and a meta-analytic test'. *Review of Educational Research*. Advance online publication. [CrossRef]

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Van Herwegen, J. and Anders, J. (2021) *What Are the Characteristics of Teacher Professional Development that Increase Pupil Achievement? A systematic review and meta-analysis*. Education Endowment Foundation. Accessed 15 April 2025. https://educationendowmentfoundation.org.uk/education-evidence/evidence-reviews/teacher-professional-development-characteristics.

Snow, P.C., Eadie, P.A., Connell, J., Dalheim, B., McCusker, H.J. and Munro, J.K. (2014) 'Oral language supports early literacy: A pilot cluster randomized trial in disadvantaged schools'. *Advances in Speech Language Pathology*, *16* (5), 495–506. [CrossRef]

Stansfield, C. (2025) 'What search terms support identification of randomised controlled trials within education research?'. OSF. Accessed 8 April 2025. http://osf.io/x875b.

Stansfield, C., Stokes, G. and Thomas, J. (2022) 'Applying machine classifiers to update searches: Analysis from two case studies'. *Research Synthesis Methods*, *13* (1), 121–33. [CrossRef]

Thomas, J., Graziosi, S., Brunton, J., Ghouze, Z., O'Driscoll, P., Bond, M. and Koryakina, A. (2022). *EPPI Reviewer: Advanced Software for systematic reviews, maps and evidence synthesis*. London: EPPI Centre, UCL Social Research Institute, University College London.

Turner, H., Boruch, R., Petrosino, A., de Moya, D., Lavenberg, J. and Rothstein, H. (2003) 'Populating an international register of randomized trials in the social, behavioral, criminological, and education sciences'. *Annals of the American Academy of Political and Social Sciences*, *589*, 3–23. [CrossRef]

Wade, A.C., Turner, H.M., Rothstein, H.R. and Lavenberg, J.G. (2006) 'Information retrieval and the role of the information specialist in producing high quality systematic reviews in the social, behavioural and education sciences'. *Evidence and Policy*, *2* (1), 89–108. [CrossRef]

Wollscheid, S. and Tripney, J. (2021) 'Rapid reviews as an emerging approach to evidence synthesis in education'. *London Review of Education*, *19* (1), 1–17. [CrossRef]