Educational Psychology Review (2024) 36:19
https://doi.org/10.1007/s10648-024-09862-5

REVIEW ARTICLE

Screening Smarter, Not Harder: A Comparative Analysis of Machine Learning Screening Algorithms and Heuristic Stopping Criteria for Systematic Reviews in Educational Research

Diego G. Campos[1] · Tim Fütterer[2] · Thomas Gfrörer[2] · Rosa Lavelle-Hill[2,3] · Kou Murayama[2] · Lars König[4] · Martin Hecht[4] · Steffen Zitzmann[2] · Ronny Scherer[1]

Accepted: 29 January 2024

Abstract

Systematic reviews and meta-analyses are crucial for advancing research, yet they are time-consuming and resource-demanding. Although machine learning and natural language processing algorithms may reduce this time and these resources, their performance has not been tested in education and educational psychology, and there is a lack of clear information on when researchers should stop the reviewing process. In this study, we conducted a retrospective screening simulation using 27 systematic reviews in education and educational psychology. We evaluated the sensitivity, specificity, and estimated time savings of several learning algorithms and heuristic stopping criteria. The results showed, on average, a 58% (SD = 19%) reduction in the screening workload of irrelevant records when using learning algorithms for abstract screening and an estimated time savings of 1.66 days (SD = 1.80). The learning algorithm random forests with sentence bidirectional encoder representations from transformers outperformed other algorithms. This finding emphasizes the importance of incorporating semantic and contextual information during feature extraction and modeling in the screening process. Furthermore, we found that 95% of all relevant abstracts within a given dataset can be retrieved using heuristic stopping rules. Specifically, an approach that stops the screening process after classifying 20% of records and consecutively classifying 5% of irrelevant papers yielded the most significant gains in terms of specificity (M = 42%, SD = 28%). However, the performance of the heuristic stopping criteria depended on the learning algorithm used and the length and proportion of relevant papers in an abstract collection. Our study provides empirical evidence on the performance of machine learning screening algorithms for abstract screening in systematic reviews in education and educational psychology.

Extended author information available on the last page of the article

Published online: 08 February 2024

Keywords Artificial intelligence · Abstract screening · Machine learning · Stopping criteria · Systematic reviews

Introduction

Systematic reviews and meta-analyses are key methods in educational research for advancing policy, research, and practice (e.g., Schneider & Preckel, 2017; van de Schoot et al., 2021; van Huizen & Plantenga, 2018). For instance, they are crucial in enabling practitioners to map the effectiveness of teaching and learning approaches, and they generate evidence to inform policy for designing evidence-based educational systems (Taylor & Hedges, 2023). However, conducting systematic reviews and meta-analyses is time-consuming and resource-intensive, with some studies estimating up to 72 weeks of skilled labor in the health sciences (Smith et al., 2011). Thus, systematic reviews that only rely on human resources are neither efficient nor sustainable, particularly with the rapid increase of scientific information (Shemilt et al., 2016).

Screening titles and abstracts is a critical component of the review process (Chai et al., 2021). During this stage, reviewers evaluate the relevance of studies identified through the initial search to determine their potential inclusion in the systematic review. Each study is typically reviewed within 30 s (Gates et al., 2018), and, given the increasing number of studies, the abstract screening phase in a systematic review requires a substantial investment of time and effort. Hence, more and more researchers are turning to learning algorithms to automate this screening phase (e.g., Guan et al., 2023; Huang et al., 2022; Scherer & Campos, 2022; Zhang et al., 2023). Prioritizing and classifying research abstracts, these algorithms efficiently identify studies that warrant further evaluation.

Studies in health science have shown that using learning algorithms can save more than 90% of the time required for manual abstract screening (van de Schoot et al., 2021). However, there is no evidence of the performance of learning algorithms in systematic reviews in education and educational psychology yet. In the health sciences, target constructs (i.e., diseases and symptoms) are usually standardized and defined by classification systems (e.g., the International Statistical Classification of Diseases and Related Health Problems; World Health Organization, 2019). In contrast, the conceptual clarity of constructs commonly used in education and educational psychology has been recently criticized (Bringmann et al., 2022; Flake & Fried, 2020; Marsh et al., 2019). Different terms refer to similar constructs, even though they have the same meaning (e.g., see Marsh et al., 2019). This lack of clarity might affect the performance of learning algorithms, particularly those trained by text data. As a result, it is necessary to gather empirical evidence about the performance of learning algorithms, the criteria for deciding when to stop reviewing, and how database characteristics may influence the accuracy of these tools in systematic reviews.

In this study, we evaluated the performance of learning algorithms and heuristic stopping rules, using a set of databases from systematic reviews in education and educational psychology. We conducted a retrospective screening simulation study

using the *ASReview* tool to determine the *sensitivity* (proportion of relevant abstracts identified), *specificity* (proportion of irrelevant abstracts that do not require review), and *estimated time savings* (reduction in time required for screening) of different learning algorithms and heuristic stopping rules.

Machine Learning for Abstract Screening

Machine learning (ML)-based tools can accelerate abstract screening in systematic reviews and meta-analyses. These tools use learning algorithms to acquire knowledge from training data, enabling the identification of specific keywords, phrases, or patterns in newly encountered abstracts and the subsequent prediction of their relevance (Marshall & Wallace, 2019). ML-based tools frequently use active learning, a form of semi-supervised learning. Active learning involves a continuous feedback loop between the researcher and the learning algorithm. Initially, the researcher provides input on which abstracts to include or exclude, and the algorithm is trained accordingly. The algorithm then suggests additional studies for the human reviewer to classify, and the process is repeated to refine the algorithm over multiple iterations. Active learning enables algorithms to learn from human feedback and select studies for labeling efficiently. At the same time, researchers have complete control over the decision to include or exclude an abstract and stop screening.

ML-based abstract screening tools involve two main stages: feature engineering and model training. Feature engineering consists of creating new variables from raw text data to transform them into quantitative formats that can be used as predictors in machine learning models. In contrast, model training involves fitting learning algorithms to the engineered dataset (Wang et al., 2022). In feature engineering, researchers must determine how to preprocess the data and generate the necessary features. Meanwhile, during model training, researchers must make critical decisions about which algorithms to use, how to train them, and how to evaluate their performance (Wang et al., 2022). In the following sections, we provide an overview of the critical elements of ML-based abstract screening tools—feature extraction methodology, model training, and stopping rule—focusing on ASReview software.

Feature Extraction

The first major stage in ML-based tools for abstract screening is feature extraction. Feature extraction is a fundamental process in natural language processing (NLP) that involves transforming raw text data into representative variables that can be used as input for the learning algorithms (Sammons et al., 2016). Effective feature extraction is essential for ML-based abstract screening tools to capture relevant information that can be used to predict the relevance of an abstract. Various methodologies have been proposed for feature extraction in NLP, including bag-of-words (BoW) (Harris, 1954), term frequency-inverse document frequency (TF-IDF) (Salton & Buckley, 1988), word embeddings (Doc2Vec) (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013), and bidirectional encoder representations from

transformers (BERT) (Reimers & Gurevych, 2019). BoW and TF-IDF are conventional approaches representing text data as a vector of word or term frequencies, respectively (Harris, 1954; Salton & Buckley, 1988). In contrast, word embeddings capture the semantic meaning of words by representing them as dense vectors in a continuous space, trained on large text corpora (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013). BERT utilizes a transformer architecture to capture bidirectional context and representations of text data, allowing for a more accurate and contextual feature extraction (Reimers & Gurevych, 2019). The selection and implementation of feature extraction methods depend on the task, dataset, and desired representation level, significantly affecting the model performance and interpretability. Evaluating feature extraction methods is critical to improving the accuracy and effectiveness of ML-based abstract screening tools.

Model Training

The second step in abstract classification involves building a predictive model. In this step, a classifier is chosen to create a model that predicts the probability of an abstract being relevant based on predictors obtained from the feature engineering process and the training data (Wang et al., 2022). These algorithms learn from the decisions of reviewers who classify studies as relevant or irrelevant and use these patterns to predict the relevance of new, unseen studies. Several classifiers can be used for abstract classification, such as logistic regression (Hosmer Jr et al., 2013), support vector machines (SVM; Platt, 1999), naive Bayes (NB;Gomes et al., 2017 ; Jackson & Moulinier, 2007), and random forests (RF; Breiman, 2017). LR is a popular linear method for binary outcome prediction, using maximum likelihood estimation to determine the coefficients of the independent variables (Hosmer Jr et al., 2013). SVM optimizes the data positions on a hyperplane to separate labels in high-dimensional spaces, making it suitable for small datasets and those with many predictor variables (Platt, 1999). Based on the Bayes theorem, NB assumes that data features are conditionally independent given their assigned class and contribute equally to the outcome (Gomes et al., 2017; Jackson & Moulinier, 2007). Lastly, RF builds multiple decision trees on data samples, preserving variance between trees and reducing overfitting (Breiman, 2017). See Bishop (2006) for a detailed overview of these approaches.

Stopping Criteria

Finally, the last central stage of ML-based tools is to decide when to stop the screening process. Active learning methods use an iterative feedback loop in which the algorithm learns from the decisions made by the human reviewer and integrates this information to predict the next document with the highest probability of inclusion. However, the final decision regarding the inclusion or exclusion of a study and when to stop screening rests with the researcher. After a reviewer decides to include or exclude a study, the algorithm incorporates this information to update its training and predict the next study deemed most relevant for review. This process continues

until the reviewer decides to stop the review process or all the studies in the literature database are classified. However, deciding *when to stop* reviewing is challenging, as users typically do not know the number of relevant studies in their literature database. Researchers may stop screening prematurely, missing relevant studies, or continue screening unnecessarily after identifying all relevant studies (Yu & Menzies, 2019). ML-based tools for abstract screening using active learning lack natural stopping criteria and continue suggesting studies until all records are classified.

Selecting an appropriate stopping criterion for ML-assisted systematic reviews is complex. Statistical and heuristic approaches have been suggested for this purpose (Callaghan & Müller-Hansen, 2020; Cormack & Grossman, 2016; Howard et al., 2020; van Haastrecht et al., 2021; Yu & Menzies, 2019). The former estimates the stopping criteria based on the derivation of an estimate of the total number of relevant papers within an initial training set and is considered more reliable than the latter (Cormack & Grossman, 2016; van Haastrecht et al., 2021). However, its practical application is limited. In contrast, heuristic approaches, including time-based, data-driven, and mixed-data strategies, are commonly used to determine when to stop reviewing (e.g., Guan et al., 2023; Huang et al., 2022; Scherer & Campos, 2022; Zhang et al., 2023). Time-based approaches involve stopping the review after reviewing a certain number of abstracts (Wallace et al., 2010). In contrast, data-driven approaches stop after identifying a certain number of consecutive irrelevant abstracts (Ros et al., 2017). Mixed strategies combine both time-based and data-driven approaches. In a mixed-based strategy, researchers decide to stop reviewing after a certain number of abstracts have been reviewed and a certain threshold of consecutive irrelevant records have been found (Hamel et al., 2021; Yu & Menzies, 2019). The use of these heuristic approaches in ML-assisted systematic reviews in education and educational psychology research is not uncommon. For example, Guan et al. (2023) used a data-driven strategy to stop the reviewing process in a systematic review of educational data ethics. Similarly, Zhang et al. (2023) used a time-based strategy to stop screening in a systematic review of school-based mental health interventions. Despite their simplicity and popularity, the fraction of relevant documents that would be retrieved using heuristic stopping criteria in education and educational psychology research is unknown.

Empirical Insights into Learning Algorithms and Heuristic Stopping Criteria

The literature on the effectiveness of various learning algorithms and heuristic stopping criteria is limited. Whereas several studies have demonstrated the potential of ML-based tools to alleviate reviewer workload and save time in the screening process (Chai et al., 2021; Gates et al., 2019; Hamel et al., 2020; Howard et al., 2020; van de Schoot et al., 2021), there is a noticeable lack of research comparing the performance of different learning algorithms (Burgard & Bittermann, 2023). Moreover, previous research into the performance of various ML-based tools for abstract screening has been limited in sample size, and the software tools evaluated were not open source, thereby hindering the identification of relevant parameters for these methods (Gates et al., 2019; Robledo et al., 2023). An attempt to

address the previous research gaps was presented by Ferdinands et al. (2020). The authors conducted a study that evaluated the performance of learning algorithms using four classifiers (LR, SVM, NB, and RF) and two feature extraction techniques (TF-IDF and D2V) on a set of six labeled datasets from systematic reviews using the software ASReview. Their findings suggested that the NB + TF-IDF learning algorithm showed the highest estimated time savings and work savings oversampling, followed by the LR + TF-IDF, SVM + TF-IDF, and RF + TF-IDF learning algorithms. However, this study had a limited sample size and did not explore the correlation between model performance and data characteristics, such as the proportion of relevant publications. In addition, the generalizability of their results to the education and educational psychology fields is uncertain. Unlike the biomedical sciences, where constructs and labels are relatively well defined, the field of education has a higher degree of ambiguity and complexity in its concepts that can affect the performance of these learning algorithms. To fully understand how data features relate to model performance and to determine the best settings for ML-based systematic reviews in education and educational psychology, it is crucial to evaluate these learning algorithms on a broader set of systematic reviews from the education and educational psychology research fields.

The performance of heuristic stopping criteria is a second aspect of ML-based tools with scarce and mixed evidence. For example, Wallace et al. (2010) found that using a time-based approach in which reviewers classified 50% of all records in a database resulted in a 100% recall. In contrast, Callaghan and Müller-Hansen (2020) found that data-driven approaches could not find 95% of all relevant records in 39% of the 20 systematic reviews analyzed. The results of these simulation studies represent an important source of information on the reliability of heuristic stopping criteria. However, a systematic evaluation of these stopping rules is still missing. Given the widespread use of heuristic stopping criteria in educational and educational psychology research, it is essential to systematically evaluate the reliability and accuracy of these stopping criteria under controlled conditions and understand how data characteristics may affect the performance of these rules. The present work aims to inform this discussion by evaluating the performance of heuristic stopping rules using a large set of systematic reviews from the fields of education and educational psychology and several learning algorithms.

ASReview as a Machine Learning Tool for Abstract Screening

There are various machine learning screening tools that are currently available (Burgard & Bittermann, 2023; van de Schoot et al., 2021). However, these tools have certain limitations. First, many of them are closed-source applications that use black-box algorithms. Second, existing tools lack the flexibility to handle the wide range of classifiers and feature extraction techniques that can be implemented in ML screening tools effectively(van de Schoot et al., 2021). In contrast, the ASReview software, an open-source platform, provides multiple learning algorithms for both feature extraction and classification. The ASReview software has gained traction among a growing number of researchers, and its potential

to reduce screening workload has been demonstrated in previous retrospective screening simulation studies (van de Schoot et al., 2021).

To incorporate ASReview software (https://asreview.nl/) into the research pipeline of a systematic review, researchers must first install the software. Once installed, the program can be operated via a web browser's graphical user interface (GUI). To initiate the abstract screening process, a project folder is created in the ASReview GUI, and the study database, including at least the titles and abstracts of the searched studies, must be added. The study database can be stored in different data formats, such as .csv, .txt, or .xlsx. ASReview runs locally on the computer, so the information is not shared with others or stored on external servers. After adding the study database, ASReview requests the user to classify some included studies as relevant or irrelevant. This subset is utilized for training an algorithm that learns the relationship between the textual features of the reviewed studies that were judged to be included or excluded. Although only a few studies must be classified, they should represent the relevant and irrelevant studies well. Following the classification, the user selects the feature extraction and classifier they want to apply. While ASReview uses a naïve Bayes classifier by default, different classifiers are available (i.e., logistic regression, random forest, and support vector machine).

The ASReview software also allows users to choose a balancing and querying strategy. Balancing strategies are employed to mitigate the risk of the learning algorithm over-fitting irrelevant studies. The default method, dynamic resampling (DR), rebalances the training set by undersampling irrelevant studies and oversampling relevant records (Ferdinands et al., 2020). Following the selection of the balancing strategy, users can proceed to choose the query strategy. The query strategy dictates the order in which studies are presented after training. The default approach, certainty-based sampling, prioritizes unscreened studies based on their predicted relevance, presenting the most relevant studies first. Alternative query strategies such as mixed, random, and uncertainty-based sampling can also be selected.

After the configuration and training phase, the learning algorithm makes a relevance prediction of all studies in the literature database and finds the study that is predicted to be most relevant. The abstract and title of the study that was predicted to be most relevant are then presented to the user, who must decide whether the study is relevant. In this way, rather than providing relevant or irrelevant information for all studies, ASReview uses a sorting mechanism based on prior inclusion and exclusion decisions to present the most likely relevant study to reviewers for the final decision. When the user has decided, this information is incorporated into the learning algorithm, and the next most relevant study is predicted. This process continues until the user decides to stop the review process or all the studies in the study database are classified. However, deciding *when to stop* reviewing is challenging, as users typically do not know the number of relevant studies in their study database. Researchers may terminate their review process too early, excluding relevant studies, or needlessly continue their review after finding all relevant studies (Yu & Menzies, 2019). ASReview currently lacks clear guidelines for when to stop, and it continues to suggest studies until all records have been reviewed.

The Present Study

ML-based tools have great potential to support the literature screening process of systematic reviews and meta-analyses. Simulation studies examining the performance of ML tools across diverse research domains indicate that these tools can reduce time requirements during the screening process (Burgard & Bittermann, 2023). However, these tools are still evolving, and their performance may vary across domains and even within the same domain (Burgard & Bittermann, 2023; Chai et al., 2021; van de Schoot et al., 2021). Additionally, the current literature lacks comprehensive comparative studies on the performance of learning algorithms when applied to real-world data (Burgard & Bittermann, 2023), and the impact of data characteristics on model performance is poorly understood (Ferdinands et al., 2020). Lastly, researchers have an ongoing uncertainty regarding the optimal number of articles to screen using these tools, and the existing simulation studies have yielded limited and conflicting findings (Callaghan & Müller-Hansen, 2020; Wallace et al., 2010).

The present study assesses the performance of various learning algorithms and heuristic stopping criteria in abstract screening within the educational and educational psychology domain. To achieve this goal, we conduct a retrospective screening simulation to compare the performance of ML screening algorithms on educational datasets and determine their sensitivity, specificity, and estimated time savings. Specifically, we address the following two research questions (RQs):

RQ1. To what extent do learning algorithms aid in saving time (*estimated time savings*) by reducing the need to screen irrelevant studies (*specificity*) in an abstract collection after finding 95% of the relevant studies for full-text screening in systematic reviews in education and educational psychology?

RQ2. How many relevant (*sensitivity*) and irrelevant (*specificity*) studies can be identified, and how much time can be saved (*estimated time savings*) in a systematic review in education and educational psychology when using learning algorithms with time-based, data-driven, and mixed-heuristic stopping criteria?

Methodology

Data Collection

We used a multistep approach to identify and collect relevant abstract collections from systematic reviews in education and educational psychology. First, we sought systematic reviews and meta-analyses in high-impact education and educational psychology journals. Second, we contacted experts in the field to inquire about any relevant studies that may have yet to be found through the initial search. We then contacted study authors of the systematic reviews and meta-analyses and requested the abstract collections they used in their studies during the abstract

screening phase. We sent 316 data requests and obtained abstract collections from 40 studies, 27 of which met the eligibility criteria. For an abstract collection to be eligible for our study, the collection must at least include information on the titles, abstracts, and screening decisions of the identified studies. For a detailed description of the data collection procedure, please see Supplementary Material S1.

In total, we included 27 databases of research syntheses in education and educational psychology. On average, the datasets included 2,738 studies (*SD* = 2,382), and 18.75% (*SD* = 15.89%) of studies were selected for full-text screening. Most of the research syntheses were made available between 2021 and 2023 (85.2%), including pre-prints and pre-publications, and the remaining were published in 2020 or earlier (14.8%). About 44.4% of the datasets were from systematic reviews, while the remaining 55.6% were from systematic reviews with meta-analyses. The research syntheses covered various topics, including factors affecting student learning (29.6%); teacher training, competence, and attitudes (22.2%); educational policies and interventions (22.2%); instructional design (14.8%); and bibliographic analyses of research methods and data (11.1%). They also targeted a wide range of research methods, such as mixed methods (48.2%), correlational studies (40.7%), and experimental studies (11.1%). The main samples of the study were K-12 students (42.3%), teachers (23.1 %), P-20 (15.4%), researchers (11.5%), and university students (7.7%). Table 1 summarizes the datasets.

Data Simulation

The abstract collections contained detailed information on the title, abstract, and screening decision (included/excluded) for each study identified in the initial search by the study authors of the systematic reviews. Studies with missing abstracts or screening decisions were excluded from the abstract collection using the statistical software R version 4.2.3 (R Core Team, 2023). The resulting abstract collections were then imported into the ASReview software to simulate the abstract review process using the simulation mode (ASReview LAB, 2023). The simulation started by selecting a learning algorithm. After selecting the learning algorithm, one included study and one excluded study were randomly selected from the abstract collection based on the screening decisions provided by the authors of the systematic review. The selection of the included and excluded studies was made at random to represent a scenario where the reviewer has minimal prior knowledge of the relevant publications and to mitigate bias in the initial training of the algorithms (Boetje & van de Schoot, in press). The learning algorithm was then trained using information obtained from these two studies, and a new study was suggested for classification. Once a new study was presented, we retrieved the classification decision (included/excluded) assigned by the authors of the systematic review. The learning algorithm was then retrained using the updated classification decision, and the process was repeated iteratively until all studies in the abstract collection were classified. It is important to note that in this retrospective simulation, rather than relying on reviewers to make the classification decision (included/excluded), we retrieved the

Table 1 Databases included in the simulation study

Database	Research topic	Total references	Percentage references selected for full-text screening	Specificity@95% (SD)
Anmarkrud et al. (2022)	Individual differences in document literacy	2502	8.5%	75% (4%)
Backfisch et al. (2020)	Measurement validity of teachers' self-reports scales	2281	31.2%	52% (2%)
Capparozza et al. (2023)	Design, evaluation, and effects of online teaching training	2613	6.5%	54% (7%)
Endedijk et al. (2022)	Teacher-student relationships and peer relationships	4210	8.5%	70% (3%)
Filges et al. (2022)	Effects of service learning on students' outcomes	7646	4.1%	89% (3%)
Filges et al. (2018)	Associations between class size and academic achievement	6216	9.8%	47% (7%)
Fong et al. (2021)	Associations between learning and study strategies	2212	8.7%	58% (6%)
Fütterer et al. (2023)	Teachers' self-assessment ability	4531	2.1%	65% (12%)
Jaeger-Dengler-Harles et al. (2020)	Information behavior in educational systematic reviews	5424	20.2%	71% (8%)
Kupers et al. (2019)	Children's creativity models	3093	31.1%	38% (8%)
Lesperance et al. (2022)	Motivational interventions and academic gender gaps	8446	1.8%	77% (5%)
Neri and Retelsdorf (2022)	Linguistic item characteristics and academic performance	686	16,8%	38% (8%)
Noetel et al. (2022)	Learning and cognitive load effects of multimedia design	1153	5.8%	56% (12%)
Pico and Woods (2023)	Effects of shared book reading on Spanish bilinguals	30	46.7%	26% (9%)
Roberts et al. (2022)	Reading instruction and reading outcomes	5063	3.3%	72% (4%)
Rowan et al. (2021)	Teacher education and student diversity	415	50.1%	35% (5%)
Saqr et al. (2022)	Network analysis in educational research	4428	39.5%	63% (6%)
Schroeder and Kucera (2022)	Effectiveness of refutation text structure in learning	461	26.9%	76% (6%)
Tarantino et al. (2022)	Teachers' attitudes and special needs education	625	5.6%	86% (7%)
Täschner et al. (2023)	Effectiveness of interventions on teacher self-efficacy	4082	19.5%	48% (7%)
Theobald (2021)	Effects of self-regulated learning on university students	518	46.3%	28% (8%)
Turan and De Smedt (2022)	Mathematical language and skills in preschoolers	131	8.4%	76% (6%)
Veletić et al. (2023)	Uses of TALIS data in the academic literature	955	46.1%	51% (5%)
Wagner et al. (2023)	Computer-based feedback and student writing	2168	2.3%	79% (6%)

Table 1 (continued)

Database	Research topic	Total references	Percentage references selected for full-text screening	Specificity@95% (SD)
Xu et al. (2022)	International students' experiences in higher education	3237	8.7%	60% (8%)
Ziernwald et al. (2023)	Reading strategies in digital reading	346	28.9%	27% (7%)
Zinsser et al. (2022)	Early childhood exclusion	467	18.8%	65% (8%)

Duplicate records and records with missing abstracts were excluded from the databases. Specificity@95%: average database specificity at 95% sensitivity

screening decisions assigned by the authors of the systematic review. By following this procedure, we could determine the number of included studies identified by the learning algorithms at different time points, thus allowing us to collect the required data to evaluate the algorithms' performance and the heuristic stopping criteria.

The simulation used ten learning algorithms for the 27 abstract collections, resulting in 270 simulation runs. Each learning algorithm comprised a feature extraction strategy and a classifier (see Table 2). To identify all relevant records in a database, we selected the default balancing strategy, "dynamic resampling," and the default query strategy, "certainty-based sampling," throughout our simulations. The initial training set of one included and one excluded study was kept constant across simulation conditions within each abstract collection. The R software version 4.2.3 (R Core Team, 2023) was used to prepare and analyze the data, and the ASReview Makita template extension was used to run the simulation study (Teijema et al., 2022). The analytic scripts can be accessed at https://osf.io/uyb7x/?view_only=74640e4eb7d146f68eb5c7738823ac9f.

Data Analysis

The analysis of the data was divided into two parts. In the initial set of analyses, we evaluated whether the learning algorithms can accelerate the identification of studies selected for full-text screening in a systematic review (RQ1). In the second set of analyses, we evaluated how many relevant studies selected for full-text screening would be identified when using learning algorithms with heuristic stopping criteria (RQ2). For the first set of analyses, we estimated the specificity and estimated time savings (ETS) at a 95% sensitivity for each learning algorithm across the 27 abstract collections. For the second set of analyses, we estimated the sensitivity, specificity, and estimated time savings of the time-based, data-driven, and mixed-strategy heuristic stopping criteria across the 27 abstract collections and ten learning algorithms. The performance metrics were defined as follows:

Table 2 Description of classifiers, feature extraction techniques, and models used in the simulation study

Feature extraction technique	Classifier	Learning algorithms
Doc2Vec	Naïve Bayes (NB)	LR+Doc2Vec
Sentence BERT (SBERT)	Logistic regression (LR)	LR+SBERT
Term frequency-inverse abstract frequency (TF-IDF)	Random forest (RF)	LR+TFIDF
	Support vector machine (SVM)	NB+TF-IDF
		RF+Doc2Vec
		RF+SBERT
		RF+TF-IDF
		SVM+Doc2Vec
		SVM+SBERT
		SVM+TF-IDF

a. Sensitivity (recall): The proportion of relevant studies in an abstract collection identified by the learning algorithms out of the total deemed relevant by the authors of the systematic review (Howard et al., 2020).
b. Specificity (true negative rate): The proportion of studies in an abstract collection that would not require screening by reviewers when using learning algorithms out of the total deemed irrelevant by the authors of the systematic review (Kusa et al., 2023).
c. Estimated time savings (ETS): The reduction in working days achieved when using learning algorithms by not having to screen irrelevant abstracts (Gates et al., 2019).

To evaluate the effectiveness of the learning algorithms (RQ1), we measured their specificity and estimated time savings while maintaining a sensitivity level of 95%. Specifically, we calculated these metrics when the learning algorithm identified 95% of the relevant studies in each abstract collection. Similarly, we evaluated the performance of the time-based, data-driven, and mixed-strategy heuristic stopping criteria (RQ2) by estimating the sensitivity, specificity, and estimate time savings the learning algorithms achieved when using these heuristic stopping criteria. The time-based stopping strategy was operationalized by dividing the total number of studies in an abstract collection into 10% screening intervals. For example, we estimated the number of relevant studies identified by the learning algorithm after screening 10% of all the studies included in an abstract collection. The data-driven strategy was operationalized by implementing a stopping criterion based on classifying consecutive irrelevant studies in an abstract collection. As the datasets had different lengths, we estimated the number of irrelevant records in relation to the total number of records in a dataset with parameters between 1 and 10%. For example, in a dataset containing 1000 studies, a data-driven stopping criterion of 1% would result in a reviewer stopping the review process after encountering ten consecutive irrelevant studies. Finally, the mixed strategy combined the parameters of the time-based and data-driven strategies, allowing reviewers to stop screening after classifying a certain percentage of studies and a predetermined threshold of consecutive irrelevant records (see Table 6).

Results

Active Learning Model Evaluation

To identify the extent to which learning algorithms can potentially save screeners time compared to random screening, we computed the average specificity and estimated time savings of the learning algorithms when they identified 95% of relevant abstracts in the abstract collections (see Table 3). The model with the highest specificity and estimated time savings was the LR+SBERT model. Using this model, a reviewer could, on average, locate 95% of all relevant studies in an abstract collection without having to screen 65% (SD = 18%) of the irrelevant abstracts. This resulted in an average estimated time savings of 1.80 days (SD = 1.94). In contrast,

Table 3 Descriptive statistics of model performance across datasets

Learning algorithms	Specificity@95%			ETS@95%		
	M	*Mdn*	*SD*	*M*	*Mdn*	*SD*
LR+Doc2Vec	57%	55%	18%	1.62	1.23	1.81
LR+SBERT	65%	71%	18%	1.80	1.29	1.94
LR+TF-IDF	61%	64%	20%	1.75	1.28	1.89
NB+TF-IDF	57%	58%	21%	1.59	1.08	1.77
RF+Doc2Vec	58%	57%	18%	1.65	1.21	1.81
RF+SBERT	63%	68%	19%	1.77	1.39	1.89
RF+TF-IDF	54%	54%	19%	1.54	1.16	1.73
SVM+Doc2Vec	53%	52%	18%	1.57	1.11	1.83
SVM+SBERT	60%	61%	19%	1.71	0.99	1.89
SVM+TF-IDF	57%	58%	20%	1.63	1.22	1.79

Specificity@95%: model specificity at 95% sensitivity. ETS@95%: estimated time-saving at 95% sensitivity.

the SVM+Doc2Vec model had the lowest specificity value of 53% (*SD* = 18%) and an average estimated time savings of 1.57 days (*SD* = 1.83). Figure 1 shows the frequency with which a model produced the highest or lowest specificity values within each of the 27 abstract collections at a 95% sensitivity. The results indicate that the RF+TF-IDF had a higher frequency of producing lower specificity values within each abstract collection. This indicates that the model required screening a higher percentage of irrelevant abstracts to locate 95% of the relevant abstracts in

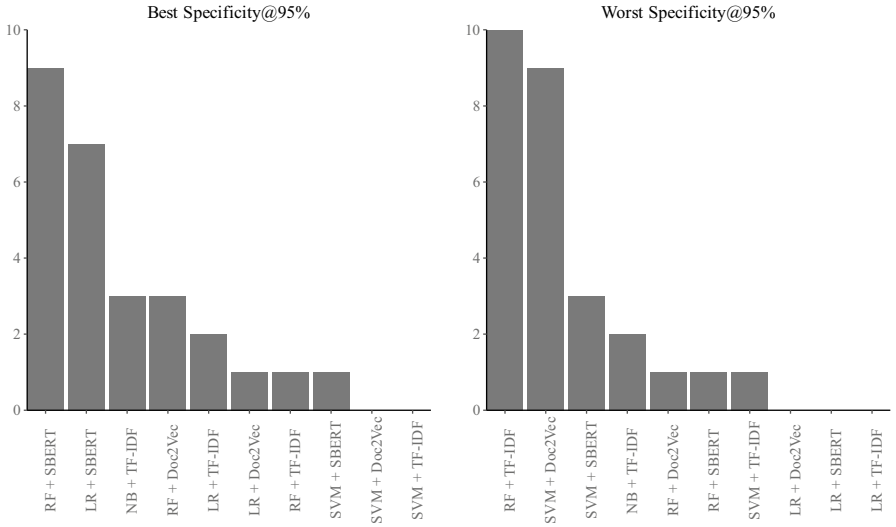

Fig. 1 Model performance based on specificity at a 95% sensitivity: top performers vs. bottom performers. Note. The figure shows the number of times a model produced the highest or lowest specificity at a 95% sensitivity within each of the 27 abstract collections

the collections. Conversely, the RF+SBERT model had a higher frequency of producing the highest specificity values. This means that using this model for abstract screening requires screening fewer irrelevant abstracts to find 95% of the relevant abstracts in an abstract collection.

The results suggest that for the sample of abstract collections from systematic reviews in education and educational psychology used in this simulation study, using active learning algorithms through ASReview could reduce the workload in identifying relevant studies for full-text screening. Unlike random screening, screening with learning algorithms does not require classifying 100% of the abstracts in an abstract collection to identify relevant abstracts (see also Supplementary Material S2). Exploratory analyses suggested a positive correlation between database length and model specificity ($\rho = .35$, 95% *CI* [0.22, 0.48]) and a negative correlation between the rate of relevant studies in an abstract collection and the specificity of the algorithms ($\rho = -0.65$, 95% *CI* [− 0.72, − 0.58]).

Evaluation of Heuristic Stopping Criteria

Time Strategy

A time-based stopping criterion means that a reviewer using learning algorithms would stop classifying abstracts in an abstract collection after reviewing between 10 and 100% of all abstracts. Table 4 presents the average sensitivity, specificity, and estimated time savings of the time-based heuristic stopping strategy across learning algorithms and abstract collections. The results indicate that to find 95% or more of the relevant abstracts in an abstract collection, a reviewer needs to screen an average of 70% of the abstracts. Hence, if a reviewer decides to stop screening after classifying 70% of the abstracts in an abstract collection, they can save, on average, 0.86 ($SD = 0.73$) days by not having to classify 36% ($SD = 6\%$) of irrelevant abstracts present in the abstract collection. However, the results varied across algorithms. The

Table 4 Performance criteria of time-based stopping criteria

Percentage reviewed	Sensitivity			Specificity			ETS		
	M	*Mdn*	*SD*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*
10%	41%	36%	22%	94%	93%	3%	2.57	2.14	2.20
20%	62%	61%	22%	86%	85%	5%	2.28	1.9	1.95
30%	76%	79%	18%	78%	76%	6%	2	1.66	1.71
40%	85%	90%	14%	69%	65%	7%	1.71	1.43	1.46
50%	91%	95%	11%	59%	55%	7%	1.43	1.19	1.22
60%	94%	98%	7%	48%	44%	7%	1.14	0.95	0.98
70%	97%	99%	5%	36%	33%	6%	0.86	0.71	0.73
80%	98%	100%	3%	24%	22%	5%	0.57	0.48	0.49
90%	100%	100%	1%	12%	11%	2%	0.28	0.24	0.24
100%	100%	100%	0%	0%	0%	1%	0	0	0

ETS estimated time saving

LR+SBERT, LR+TF-IDF, RF+SBERT, and SVM+SBERT algorithms achieved 95% sensitivity with a stopping criterion of 60% of classified studies. Figure 2 provides a visual representation of the percentage of relevant abstracts identified by each learning algorithm as a function of the percentage of classified abstracts.

The sensitivity and specificity of the time-based stopping criterion also varied across abstract collections. In some abstract collections, it was possible to identify 95% of the relevant abstracts after classifying 20% of the abstracts in the abstract collection. In contrast, other databases were required to classify 90% of all the abstracts in the abstract collection to achieve the same 95% sensitivity. Exploratory analyses indicated a strong positive Spearman's rank correlation between the average percentage of classified abstracts and the proportion of relevant abstracts in an abstract collection ($\rho = 0.81$, 95% *CI* [0.61, 0.90]). Conversely, Spearman's rank correlation between the number of irrelevant records and the percentage of classified

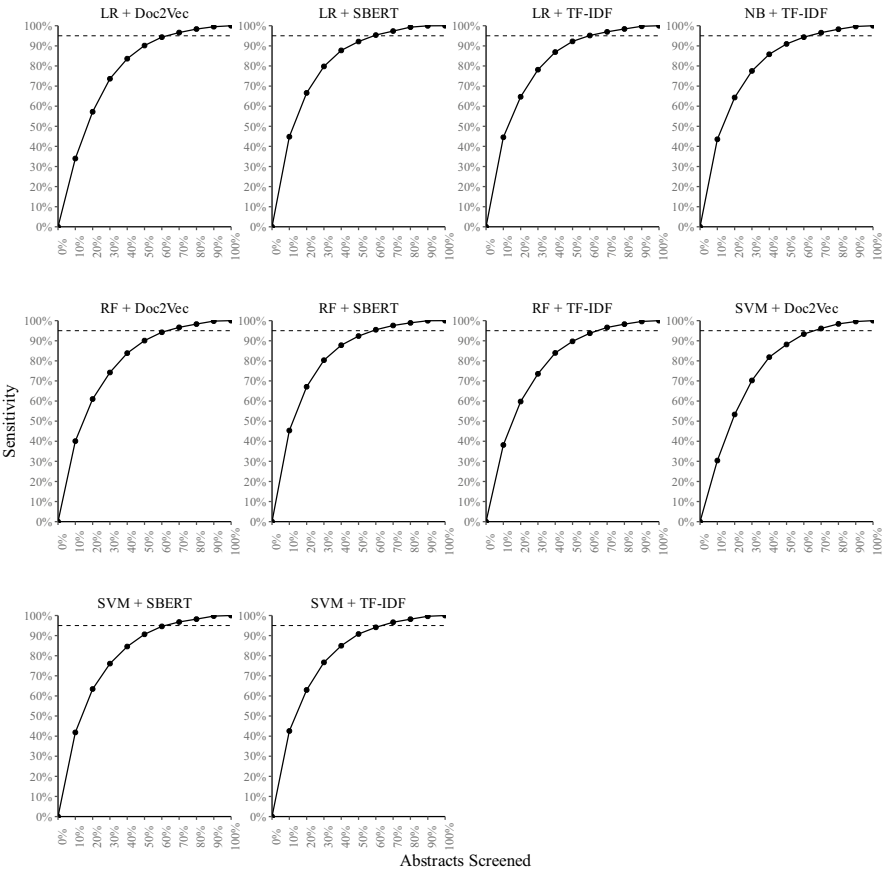

Fig. 2 Sensitivity curves of learning algorithms using a time-driven approach. Note. The figure displays the mean percentage of relevant abstracts identified by each learning algorithm after classifying 10 to 100% of all abstracts in the abstract collections

abstracts was negative ($\rho = -0.81$, 95% *CI* [− 0.91, − 0.61]). Supplementary Material S2 presents the percentage of relevant records identified in each abstract collection by different learning algorithms after classifying 10 to 100% of the abstracts in each collection.

Data-Driven Strategy

A reviewer using data-driven stopping criteria would stop the classification of abstracts after classifying between 1 and 10% consecutive irrelevant abstracts. Table 5 presents the average sensitivity, specificity, and estimated time savings of the data-driven heuristic stopping criteria across all abstract collections and learning algorithms. The results indicate that to identify 95% of the relevant abstracts in an abstract collection, a reviewer would need to classify, on average, 7% of consecutive irrelevant abstracts. This means that, on average, the reviewer would not have to classify 38% (*SD* = 27%) of the irrelevant abstracts in an abstract collection, resulting in an estimated time saving of 1.04 (*SD* = 1.47) days.

The sensitivity of the time-based stopping criteria varied across learning algorithms. Figure 3 shows the average percentage of relevant records identified by each learning algorithm as a function of the percentage of irrelevant records classified in a row. A reviewer using the RF+SBERT model would, on average, identify 95% of the relevant records in the abstract collections using a stopping criterion of classifying 6% of irrelevant studies in a row. In contrast, using the RF + Doc2Vec and SVM + Doc2Vec algorithms with a time-based stopping criterion between 1 and 10% of irrelevant abstracts classified in a row would not identify 95% of the relevant abstracts in an abstract collection.

The performance of the data-based stopping criteria also varied across abstract collections. The minimum percentage of irrelevant records classified in a row required to identify 95% of all relevant abstracts in an abstract collection was 1%,

Table 5 Performance criteria of data-driven stopping criteria

Percentage irrelevant	Sensitivity			Specificity			ETS		
	M	*Mdn*	*SD*	*M*	*Mdn*	*SD*	*M*	*Mdn*	*SD*
1%	59%	74%	36%	81%	88%	18%	2	1.32	1.98
2%	79%	93%	29%	65%	71%	25%	1.64	1	1.85
3%	87%	97%	24%	55%	58%	27%	1.43	0.79	1.75
4%	90%	98%	21%	49%	53%	27%	1.28	0.57	1.65
5%	92%	99%	19%	45%	48%	28%	1.16	0.50	1.58
6%	93%	99%	19%	41%	45%	28%	1.09	0.47	1.51
7%	95%	99%	16%	38%	43%	27%	1.04	0.45	1.47
8%	95%	100%	16%	36%	39%	26%	0.97	0.43	1.39
9%	95%	100%	16%	35%	36%	26%	0.93	0.38	1.36
10%	95%	100%	16%	33%	32%	25%	0.89	0.34	1.31

ETS estimated time saving

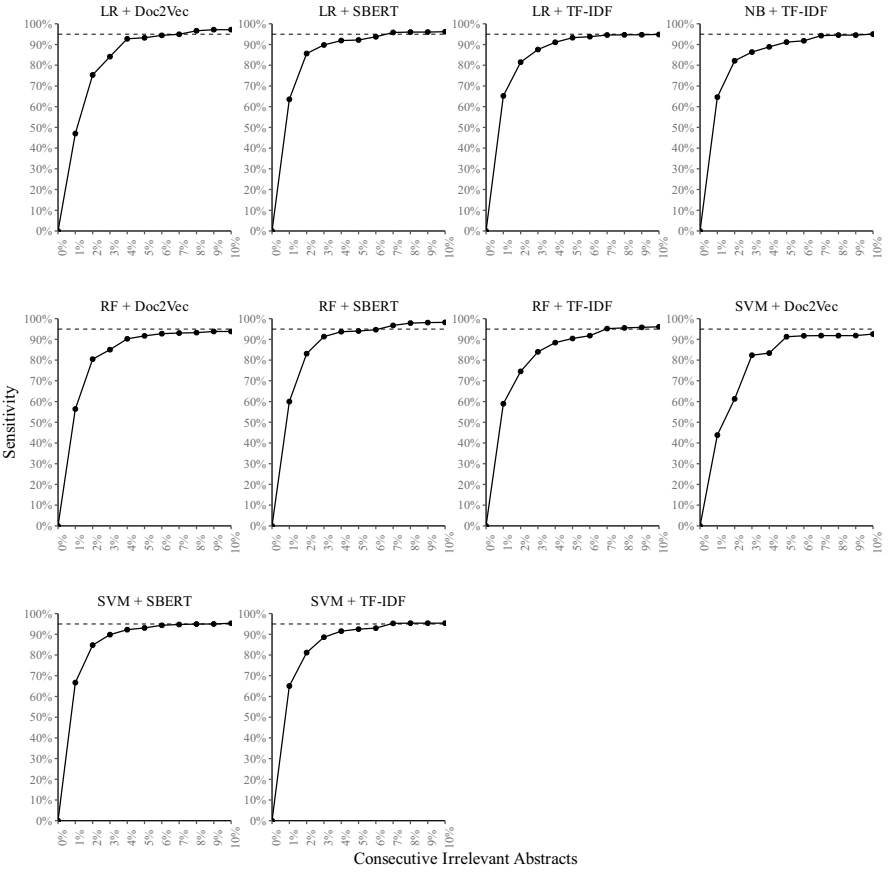

Fig. 3 Sensitivity curves of learning algorithms using a data-based approach. Note. The figure shows the average percentage of relevant abstracts identified by each learning algorithm after classifying 1 to 10% of consecutive irrelevant records

while the maximum was 9%. In two abstract collections, the stopping criteria of classifying 10% of irrelevant abstracts in a row was insufficient to achieve a 95% sensitivity. Supplementary Material S3 shows the percentage of relevant abstracts identified by each learning algorithm for each dataset when using a data-driven stopping criteria of classifying 1 to 10% of consecutive irrelevant abstracts.

Mixed Strategy

A reviewer using a mixed-based heuristic stopping strategy would decide to stop the abstract screening after classifying a certain percentage of all the abstracts included in an abstract collection and classifying a certain percentage of consecutive abstracts as irrelevant. Table 6 presents the sensitivity and specificity of the nine mixed-based heuristic stopping criteria evaluated in this study (labeled as groups A to I). On

Table 6 Performance criteria of mixed-based stopping criteria

Mixed strategy	Sensitivity			Specificity		
	M	*Mdn*	*SD*	*M*	*Mdn*	*SD*
A	71%	78%	26%	78%	85%	16%
B	93%	99%	19%	43%	48%	29%
C	96%	100%	16%	31%	31%	27%
D	78%	87%	21%	74%	80%	14%
E	95%	99%	13%	42%	48%	28%
F	97%	100%	10%	30%	31%	26%
G	85%	91%	16%	69%	72%	11%
H	96%	99%	11%	41%	48%	26%
I	98%	100%	9%	30%	31%	26%

A: 10% records screened and 1% irrelevant records. B: 10% records screened and 5% irrelevant records. C: 10% records screened and 10% irrelevant records. D: 20% records screened and 1% irrelevant records. E: 20% records screened and 5% irrelevant records. F: 20% records screened and 10% irrelevant records. G: 30% records screened and 1% irrelevant records. H: 30% records screened and 5% irrelevant records. I: 30% of records screened and 10% of irrelevant records. WSS: work saved over sampling

average, groups C, E, F, H, and I were able to identify 95% or more of the relevant abstracts in an abstract collection, with group E having the highest average specificity ($\overline{X} = 42\%$, $SD = 28\%$), indicating that on average, a reviewer would not have to screen 42% of the irrelevant records in an abstract collection to identify 95% of the relevant abstracts. The performance of the mixed-based strategies varied across algorithms, with the RF+SBERT model delivering the highest average sensitivity ($\overline{X} = 92\%$, $SD = 8\%$) and specificity ($\overline{X} = 52\%$, $SD = 19\%$) across strategies. In contrast, the SVM+Doc2Vec model displayed the lowest average sensitivity ($\overline{X} = 87\%$, $SD = 13\%$) and specificity values ($\overline{X} = 44\%$, $SD = 20\%$). Figure 4 shows the average percentage of relevant abstracts identified by the learning algorithms with mixed-based heuristic stopping criteria and the average percentage of irrelevant abstracts a reviewer would not need to classify.

Exploratory analyses using a Spearman's rank correlation indicated that the sensitivity achieved under each mixed-based heuristic stopping criteria had a positive correlation with the number of abstracts in an abstract collection in groups A, D, and G (group A: $\rho = 0.76$, 95% *CI* [0.50, 0.89]; group D: $\rho = 0.75$, 95% *CI* [0.47, 0.87]; group G: $\rho = 0.68$, 95% *CI* [0.38, 0.87]), but not in groups B, C, E, F, H, and I (group B: $\rho = 0.36$, 95% *CI* [− 0.09, 0.67]; group C: $\rho = -0.08$, 95% *CI* [− 0.55, 0.38]; group E: $\rho = 0.37$, 95% *CI* [− 0.06, 0.70]; group F: $\rho = -0.11$, 95% *CI* [− 0.55, 0.36]; group H: $\rho = 0.32$, 95% *CI* [− 0.11, 0.65]; group I: $\rho = -0.11$, 95% *CI* [− 0.56, .34]. Furthermore, we observed a positive correlation between the sensitivity achieved by the mixed-based heuristic stopping criteria and the number of relevant studies in an abstract collection in groups C, F, I, and G (group C: $\rho = 0.66$, 95% *CI* [0.25, 0.90]; group F: $\rho = 0.67$, 95% *CI* [0.28, 0.91]; group I: $\rho = 0.67$, 95% *CI* [0.25, 0.90]; and group G:

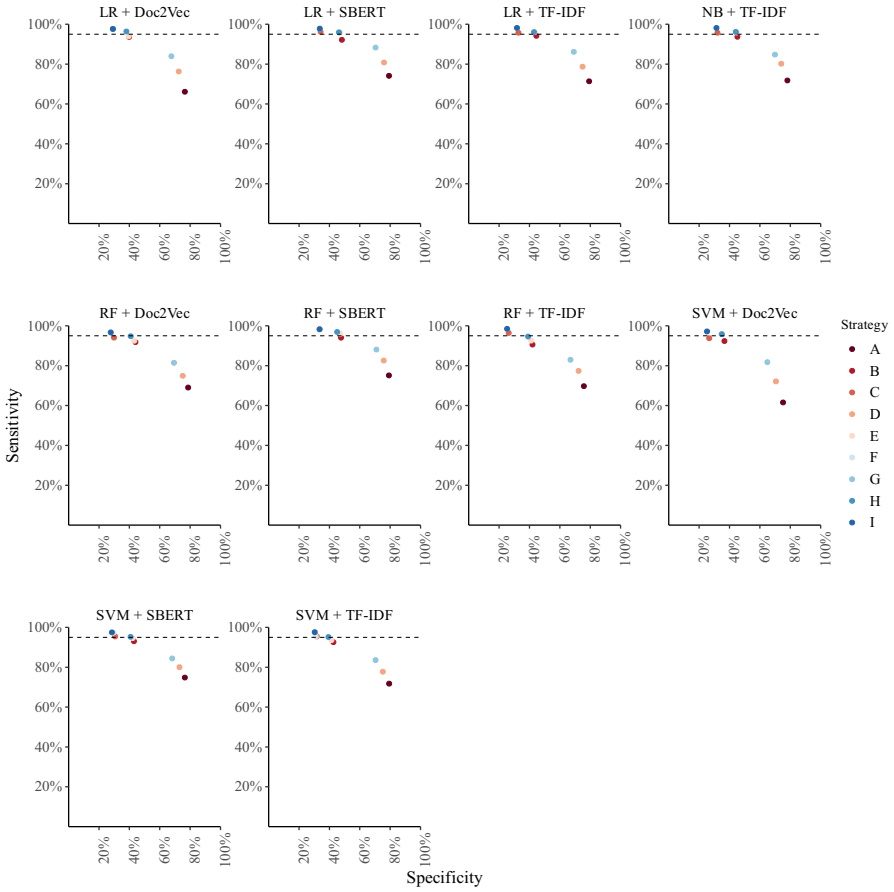

Fig. 4 Sensitivity and specificity of learning algorithms using a mixed-based approach. *Note.* A: 10% of records reviewed and 1% of irrelevant records classified in a row. B: 10% of records reviewed and 5% of irrelevant records in a row. C: 10% of records reviewed and 10% of irrelevant records classified in a row. D: 20% of records reviewed and 1% of irrelevant records classified in a row. E: 20% of records reviewed and 5% of irrelevant records classified in a row. F: 20% of records reviewed and 10% of irrelevant records classified in a row. G: 30% of records reviewed and 1% of irrelevant records classified in a row. H: 30% of records reviewed and 5% of irrelevant records classified in a row. I: 30% of records reviewed and 10% of irrelevant records classified in a row

$\rho = -0.50$, 95% *CI* [− 0.77, − 0.09]); and no correlation in groups A, B, D, E, and H (group A: $\rho = -0.11$, 95% *CI* [− 0.50, 0.31]; group B: $\rho = 0.41$, 95% *CI* [− 0.01, 0.73]; group D: $\rho = -0.24$, 95% *CI* [− 0.60, 0.19]; group E: $\rho = 0.39$, 95% *CI* [0.00, 0.72]; group H: $\rho = 0.32$, 95% *CI* [− 0.10, 0.68]). Supplementary Material S4 presents the percentage of relevant abstracts and the saved screening effort achieved using learning algorithms with mixed-based heuristic stopping criteria for each of the 27 abstract collections.

Discussion

Learning Algorithms

In this retrospective simulation study, we evaluated the sensitivity, specificity, and estimated time savings of learning algorithms in abstract collections from systematic reviews in educational and educational psychology. The results from this study suggest that active learning can reduce the effort required for abstract screening in systematic reviews. The amount of work saved in screening time varied depending on the learning algorithm and database used. The LR+SBERT model outperformed the other models in terms of specificity (M = 65%, SD = 18%) and ETS (M = 1.80, SD = 1.94) at a 95% sensitivity, highlighting the importance of incorporating semantic and contextual information during feature extraction and modeling in educational and educational psychology research. These results are similar to those presented in the systematic review by Burgard and Bittermann (2023), who reported an average work savings over sampling of 55% at a sensitivity of 95% across 21 studies. This means that, on average, screening tools could identify 95% of the relevant abstracts in an abstract collection without needing to classify 55% of the abstracts. However, they are lower than the estimates reported by van de Schoot et al. (2021), who found a mean work savings over sampling of 83% at a 95% sensitivity in systematic reviews in software engineering, psychology, and medicine. The complex inclusion criteria of the educational and educational psychology systematic reviews included in this study might contribute to the lower specificity estimates. For example, Rowan et al. (2021) systematic review of teacher education and diverse learners focused primarily on the literature on initial teacher education. This systematic review did not include studies that discussed broader issues such as education, schools, or school systems. As "initial teacher training" and "education" overlap in meaning, it is difficult for a learning algorithm to accurately differentiate between relevant and irrelevant studies, contributing to the lower specificity achieved in this study.

A second factor that may be related to the lower specificity estimates achieved in our simulation study is the substantial proportion of relevant abstracts in the systematic reviews. For example, in the Theobald (2021) review of the effects of self-regulated learning on university students, it resulted in 46.3% of records for full-text inclusion. Because there are more relevant studies in the dataset, ML-based abstract review tools require more time to find all relevant studies, increasing screening time. Thus, systematic reviews with highly specific search terms and inclusion/exclusion criteria will likely have higher specificity estimates.

Third, the lack of standardized concepts in education and educational psychology could also explain the lower specificity estimates achieved in our study. Previous research suggests that the field of education and educational psychology uses different terms to refer to constructs that are conceptually and empirically similar (e.g., see Marsh et al., 2019). This issue is prevalent in teaching research, which contains a significant number of ambiguous concepts, such as teacher competence or teaching quality (e.g., see Blömeke et al., 2015; Senden et al., 2022).

Since a significant portion of the abstract collections in our study came from this field, learning algorithms that rely on feature extraction techniques based on key concepts and phrases may have faced difficulties finding records that use different concepts but refer to the same construct. The use of complex learning algorithms (RF+SBERT or LR + SBERT) could overcome this issue as these algorithms can incorporate semantic and contextual information. Further research comparing different ML tools would be beneficial in gaining a more comprehensive understanding of the performance of these learning algorithms in educational research synthesis. Our current study provides a foundational step in this area, and we recommend it as an essential focus for future research.

Overall, the results suggest that learning algorithms can reduce screening time in a wide range of educational and educational psychology systematic reviews and point to differences between learning algorithms that should be considered when using an ML-based abstract screening tool. As the learning algorithms (e.g., RF + SBERT) investigated in this study are also used in other ML-based screening tools, the performance insights gained in this study are likely to apply to other tools employing these shared learning algorithms. The use of active learning algorithms on text data from the education field was not previously studied. However, this research demonstrates that these algorithms perform well in this field. Researchers could use learning algorithms for abstract screening in educational and educational psychology to decrease the likelihood of human error in systematic reviews. Previous research suggests that non-experienced reviewers can miss up to 13% of the records in a systematic review, which can significantly impact the results of a meta-analysis (Waffenschmidt et al., 2019). Thus, using learning algorithms for abstract screening in educational and educational psychology can decrease human error in systematic reviews, leading to more accurate and reliable outcomes in systematic reviews and meta-analyses.

Heuristic Stopping Rules

Our second research objective was to evaluate how many relevant studies can be identified and how much time can be saved in a systematic review of education and educational psychology when using learning algorithms with time-based, data-driven, and mixed-heuristic stopping criteria. The findings indicated that different heuristic stopping rules could achieve a sensitivity of 95%. For example, a time-based strategy using a stopping rule of 70% or more of studies screened may achieve a sensitivity of 95%, whereas a data-based strategy using a stopping rule of classifying 7% or more of irrelevant studies in a row may achieve a sensitivity of 95%. Moreover, in the mixed-based strategy, achieving a sensitivity of 95% was possible with various stopping rules, including screening 10% of all studies and classifying 10% of irrelevant studies in a row, screening 20% of studies and classifying 5% or 10% of irrelevant studies in a row, and screening 30% of studies and classifying 5% or 10% of irrelevant studies in a row.

Consistent with prior studies, our exploratory analysis suggests that the effectiveness of heuristic strategies depends on various factors, including the database size,

the percentage of relevant abstracts (König et al., 2023), the topic complexity, and the learning algorithm used (Howard et al., 2020). For instance, the findings indicate that a data-driven approach could not identify 95% of the relevant abstracts in abstract collections with fewer than 200 studies. Similarly, for abstract collections with more than 40% of relevant studies, the data-driven strategy required more classified abstracts to identify 95% of the relevant records. The variability in the performance of heuristic stopping criteria across abstract collections may also be related to the complexity of the topics in educational research. Heuristic stopping criteria may be less effective when handling research areas with multiple concepts referring to the same educational phenomena or when the inclusion/exclusion criteria are ambiguous. Active learning algorithms prioritize studies that are similar to previously identified relevant studies. As a result, studies that differ conceptually or methodologically from previously identified relevant studies may not be considered. Thus, the same factors that pose challenges for designing a search with adequate sensitivity and specificity may also pose problems for active learning algorithms and the use of heuristic stopping rules (Tsou et al., 2020).

Overall, the performance of heuristic stopping rules should not be generalized to databases where the number of relevant records is unknown, as their development depends on knowledge of the true relevant records (Callaghan & Müller-Hansen, 2020; Hamel et al., 2021). Researchers should be cautious when using heuristic stopping criteria in systematic reviews with a limited number of identified publications or in fields where many relevant documents are expected. Estimating the prevalence of relevant abstracts before selecting a stopping rule could help researchers to make informed decisions on when to stop screening (König et al., 2023). Future research should evaluate the performance of statistical stopping criteria in the educational and educational psychology research fields (Callaghan & Müller-Hansen, 2020; Cormack & Grossman, 2016; Howard et al., 2020; Yu & Menzies, 2019).

Limitations and Future Direction

Our study addresses the limited availability of data on the performance of ASReview in educational and educational psychology research and contributes to the ongoing discussion surrounding the use of heuristic stopping criteria in ML-based systematic reviews. It is important to note that although our sample included a heterogeneous range of systematic reviews in education and educational psychology research, with variations in topics, database length, and proportion of included studies, it was neither exhaustive nor representative of the entire population. Moreover, a considerable proportion of the systematic reviews have not yet been published and undergone peer review for quality insurance, which could impact the results reported in this study. Therefore, caution should be exercised in extrapolating these findings to the broader population of systematic reviews in education and educational psychology. A second limitation of this study relates to selecting the initial training set of relevant and irrelevant studies, which was chosen randomly and held constant across simulation conditions. To address the challenges of active learning models in identifying a diverse set of conceptual or methodologically relevant studies, researchers

could expand the training set to include a more extensive and diverse sample. This would help the learning algorithms to become more sensitive to the methodological and conceptual variability of potentially relevant studies for systematic reviews in educational psychology research. Further research is necessary to determine how the performance of learning algorithms is affected by variations in the initial training set of relevant studies. Third, our study included a comprehensive set of learning algorithms in the ASReview software. However, it is crucial to expand the analysis to include other algorithms and gain a deeper understanding of how those algorithms interact with the concepts and terminology of educational psychology research. Future studies could use the abstract collections collected for this retrospective screening simulation to understand the impact of content-specific factors in the abstract collections while controlling for sample size and prevalence. These analyses may reveal features that affect the performance of learning algorithms and heuristic strategies in the educational psychology literature. Finally, given that the results on the performance of heuristic stopping criteria are highly dependent on dataset characteristics, it is advisable to extend the study to evaluate the effectiveness and generalizability of statistical stopping criteria, such as those proposed by Callaghan and Müller-Hansen (2020).

Conclusions

This study examined various learning algorithms and evaluated the effectiveness of different heuristic stopping criteria in reducing the time and effort required for abstract screening in systematic reviews in educational psychology research. The findings indicated that active learning could expedite the screening process for education researchers by enabling them to identify relevant records faster than random screening. The time-based, data-driven, and mixed strategies were also found to achieve 95% sensitivity, making them a practical option for stopping screening. However, the choice of a heuristic stopping rule should be carefully considered, given the dependence of results on the characteristics of the databases and algorithms employed. Applied researchers should balance the need to identify all relevant records against time constraints when selecting a stopping rule.

Overall, using learning algorithms in the abstract screening phase provides multiple opportunities for educational psychology researchers. First, systematic reviews require a comprehensive identification of relevant literature. The screening phase is about efficiently finding relevant studies, rather than understanding each paper in-depth. By simplifying this initial screening phase, machine learning tools such as ASReview expedite the identification of pertinent papers, allowing researchers to dedicate more time to in-depth analysis and synthesis of the selected literature. Thus, using machine learning tools for abstract screening can enhance the quality and efficiency of systematic reviews.

Second, researchers could implement more sensitive search strategies. As learning algorithms can reduce the time, fatigue, and human error in the abstract screening phase, researchers could broaden the scope of their search strategies to identify a more comprehensive set of relevant studies in the abstract screening phase. However, researchers

should be aware that active learning algorithms rely on previous classification decisions and prioritize studies with similar study designs or conceptual frameworks to previously identified relevant studies. Therefore, there is a risk of overlooking studies that differ methodologically or conceptually from previously identified relevant studies. Improving our understanding of how learning algorithms deal with the diverse methodological and conceptual approaches in educational psychology research is critical.

Third, researchers should be critical of using abstract screening tools with *black boxes* or *out-of-the-box* solutions. The selection of a learning algorithm and stopping criteria impacts the percentage of relevant studies in the abstract screening phase. The conceptual and methodological diversity found in educational research suggests the need to fine-tune the learning algorithms and consider using feature extractors that use semantic and contextual information (SBERT) rather than frequency counts of keywords (TF-IDF). Future studies using ML-based abstract screening tools should report on the type of algorithms and stopping criteria used to improve the transparency and reproducibility of the systematic reviews. The results presented in this study guide researchers who wish to optimize the screening process in systematic reviews using the ML-based abstract screening tool ASReview.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10648-024-09862-5.

Funding Open access funding provided by University of Oslo (incl Oslo University Hospital). This research was supported in part by the LEAD Intramural Research Funding from the University of Tubingen and the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research to Kou Murayama.

Data Availability Data used in this study are available in the Open Science Framework repository https://osf.io/uyb7x/.

Declarations

Competing Interests The authors declare that they have no competing interests regarding the publication of this article. There are no financial, personal, or professional conflicts of interest that could potentially influence the interpretation or presentation of the research findings.

References

Anmarkrud, Ø., Bråten, I., Florit, E., & Mason, L. (2022). The role of individual differences in sourcing: A systematic review. *Educational Psychology Review, 34*(2), 749–792. https://doi.org/10.1007/s10648-021-09640-7

ASReview LAB. (2023). *ASReview LAB - A tool for AI-assisted systematic reviews* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.7672035

Backfisch, I., Schneider, J., Lachner, A., Scheiter, K., & Scherer, R. (2020). *Another jingle-jangle fallacy?Examining the validity of Technological Pedagogical and Content Knowledge (TPACK) self-report assessments*. https://www.psycharchives.org/en/item/50b6f757-52d3-4902-863a-d833279f3ce2

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 4, Issue 4). Springer.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies: Competence viewed as a continuum. *Zeitschrift Für Psychologie, 223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194

Breiman, L. (2017). *Classification and regression trees*. Routledge. https://doi.org/10.1201/9781315139470

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science, 31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Burgard, T., & Bittermann, A. (2023). Reducing literature screening workload with machine learning. *Zeitschrift Für Psychologie, 231*(1), 3–15. https://doi.org/10.1027/2151-2604/a000509

Callaghan, M. W., & Müller-Hansen, F. (2020). Statistical stopping criteria for automated screening in systematic reviews. *Systematic Reviews, 9*(1), 273. https://doi.org/10.1186/s13643-020-01521-4

Capparozza, M., Fröhlich, N., Dehmel, A., & Fauth, B. (2023). Gestaltung und evaluation von webbasierten Lehrkräftefortbildungen: Ein systematic review. In K. Scheiter & I. Gogolin (Eds.), *Bildung für eine digitale Zukunft* (pp. 363–397). Springer Fachmedien. https://doi.org/10.1007/978-3-658-37895-0_15

Chai, K. E. K., Lines, R. L. J., Gucciardi, D. F., & Ng, L. (2021). Research screener: A machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic Reviews, 10*(1), 93. https://doi.org/10.1186/s13643-021-01635-3

Cormack, G. V., & Grossman, M. R. (2016). Engineering quality and reliability in technology-assisted review. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 75–84). https://doi.org/10.1145/2911451.2911510

Endedijk, H. M., Breeman, L. D., van Lissa, C. J., Hendrickx, M. M. H. G., den Boer, L., & Mainhard, T. (2022). The teacher's invisible hand: A meta-analysis of the relevance of teacher–student relationship quality for peer relationships and the contribution of student behavior. *Review of Educational Research, 92*(3), 370–412. https://doi.org/10.3102/00346543211051428

Ferdinands, G., Schram, R. D., Bruin, J. de, Bagheri, A., Oberski, D. L., Tummers, L., & Schoot, R. van de. (2020). *Active learning for screening prioritization in systematic reviews—A simulation study*. OSF Preprints. https://doi.org/10.31219/osf.io/w6qbg

Filges, T., Dietrichson, J., Viinholt, B. C. A., & Dalgaard, N. T. (2022). Service learning for improving academic success in students in grade K to 12: A systematic review. *Campbell Systematic Reviews, 18*(1), e1210. https://doi.org/10.1002/cl2.1210

Filges, T., Sonne-Schmidt, C. S., & Nielsen, B. C. V. (2018). Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews, 14*(1), 1–107. https://doi.org/10.4073/csr.2018.10

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fong, C. J., Krou, M. R., Johnston-Ashton, K., Hoff, M. A., Lin, S., & Gonzales, C. (2021). LASSI's great adventure: A meta-analysis of the learning and study strategies inventory and academic outcomes. *Educational Research Review, 34*, 100407. https://doi.org/10.1016/j.edurev.2021.100407

Fütterer, T., Tschönhens, F., Scherer, R., Dickhäuser, O., & Ruiz-Primo, M. A. (2023). *Teachers' self-assessment ability—A systematic literature review and meta-analysis*. [Manuscript in preparation]. Faculty of Economic and Social Sciences. University of Tübingen.

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Systematic Reviews, 8*(1), 278. https://doi.org/10.1186/s13643-019-1222-2

Gates, A., Johnson, C., & Hartling, L. (2018). Technology-assisted title and abstract screening for systematic reviews: A retrospective evaluation of the Abstrackr machine learning tool. *Systematic Reviews, 7*(1), 45. https://doi.org/10.1186/s13643-018-0707-8

Gomes, S. R., Saroar, S. G., Mosfaiul, M., Telot, A., Khan, B. N., Chakrabarty, A., & Mostakim, M. (2017). A comparative approach to email classification using Naive Bayes classifier and hidden Markov model. In *2017 4th International Conference on Advances in Electrical Engineering (ICAEE)* (pp. 482–487).

Guan, X., Feng, X., & Islam, A. Y. M. A. (2023). The dilemma and countermeasures of educational data ethics in the age of intelligence. *Humanities and Social Sciences Communications, 10*(1), 1–14. https://doi.org/10.1057/s41599-023-01633-x

Hamel, C., Hersi, M., Kelly, S. E., Tricco, A. C., Straus, S., Wells, G., Pham, B., & Hutton, B. (2021). Guidance for using artificial intelligence for title and abstract screening while conducting knowledge syntheses. *BMC Medical Research Methodology, 21*(1), 285. https://doi.org/10.1186/s12874-021-01451-2

Hamel, C., Kelly, S. E., Thavorn, K., Rice, D. B., Wells, G. A., & Hutton, B. (2020). An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening – impact on reviewer-relevant outcomes. *BMC Medical Research Methodology, 20*(1), 256. https://doi.org/10.1186/s12874-020-01129-1

Harris, Z. S. (1954). Distributional Structure. *WORD, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Howard, B. E., Phillips, J., Tandon, A., Maharana, A., Elmore, R., Mav, D., Sedykh, A., Thayer, K., Merrick, B. A., Walker, V., Rooney, A., & Shah, R. R. (2020). SWIFT-Active screener: Accelerated document screening through active learning and integrated recall estimation. *Environment International, 138*, 105623. https://doi.org/10.1016/j.envint.2020.105623

Huang, Y., Procházková, M., Lu, J., Riad, A., & Macek, P. (2022). Family related variables' influences on adolescents' health based on health behaviour in school-aged children database, an AI-assisted scoping review, and narrative synthesis. *Frontiers in Psychology, 13* https://www.frontiersin.org/articles/10.3389/fpsyg.2022.871795

Jackson, P., & Moulinier, I. (2007). *Natural language processing for online applications: Text retrieval, extraction and categorization* (Vol. 5). John Benjamins Publishing.

Jaeger-Dengler-Harles, I., Heck, T., & Rittberger, M. (2020). Systematic reviews as object to study relevance assessment processes. In *Proceedings of ISIC, the Information Behaviour Conference, Pretoria, South Africa* (Vol. 25). Internet Archive. https://doi.org/10.47989/irisic2024

König, L., Zitzmann, S., Fütterer, T., Campos, D. G., Scherer, R., & Hecht, M. (2023). When to stop and what to expect—An evaluation of the performance of stopping rules in AI-assisted reviewing for psychological meta-analytical research. *PsyArXiv*. https://doi.org/10.31234/osf.io/ybu3w

Kupers, E., Lehmann-Wermser, A., McPherson, G., & van Geert, P. (2019). Children's creativity: A theoretical framework and systematic review. *Review of Educational Research, 89*(1), 93–124. https://doi.org/10.3102/0034654318815707

Kusa, W., Lipani, A., Knoth, P., & Hanbury, A. (2023). An analysis of work saved over sampling in the evaluation of automated citation screening in systematic literature reviews. *Intelligent Systems with Applications, 18*, 200193. https://doi.org/10.1016/j.iswa.2023.200193

Lesperance, K., Hofer, S., Retelsdorf, J., & Holzberger, D. (2022). Reducing gender differences in student motivational-affective factors: A meta-analysis of school-based interventions. *British Journal of Educational Psychology, 92*(4), 1502–1536. https://doi.org/10.1111/bjep.12512

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111*(2), 331–353. https://doi.org/10.1037/edu0000281

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews, 8*(1), 163. https://doi.org/10.1186/s13643-019-1074-9

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv. http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Neri, N., & Retelsdorf, J. (2022). The role of linguistic features in science and math comprehension and performance: A systematic review and desiderata for future research. *Educational Research Review, 36*, 100460. https://doi.org/10.1016/j.edurev.2022.100460

Noetel, M., Griffith, S., Delaney, O., Harris, N. R., Sanders, T., Parker, P., del Pozo Cruz, B., & Lonsdale, C. (2022). Multimedia design for learning: An overview of reviews with meta-meta-analysis. *Review of Educational Research, 92*(3), 413–454. https://doi.org/10.3102/00346543211052329

Pico, D. L., & Woods, C. (2023). Shared book reading for Spanish-speaking emergent bilinguals: A review of experimental studies. *Review of Educational Research, 93*(1), 103–138. https://doi.org/10.3102/00346543221095112

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers, 10*(3), 61–74.

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese bert-networks. arXiv. http://arxiv.org/abs/1908.10084

Roberts, G. J., Dumas, D. G., McNeish, D., & Coté, B. (2022). Understanding the dynamics of dosage response: A nonlinear meta-analysis of recent reading interventions. *Review of Educational Research, 92*(2), 209–248. https://doi.org/10.3102/00346543211051423

Robledo, S., Grisales Aguirre, A. M., Hughes, M., & Eggers, F. (2023). "Hasta la vista, baby" – Will machine learning terminate human literature reviews in entrepreneurship? *Journal of Small Business Management, 61*(3), 1314–1343. https://doi.org/10.1080/00472778.2021.1955125

Ros, R., Bjarnason, E., & Runeson, P. (2017). A machine learning approach for semi-automated search and selection in literature studies. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering* (pp. 118–127). https://doi.org/10.1145/3084226.3084243

Rowan, L., Bourke, T., L'Estrange, L., Lunn Brownlee, J., Ryan, M., Walker, S., & Churchward, P. (2021). How does initial teacher education research frame the challenge of preparing future teachers for student diversity in schools? A systematic review of literature. *Review of Educational Research, 91*(1), 112–158. https://doi.org/10.3102/00346543209791171

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523.

Sammons, M., Christodoulopoulos, C., Kordjamshidi, P., Khashabi, D., Srikumar, V., Vijayakumar, P., Bokhari, M., Wu, X., & Roth, D. (2016). EDISON: Feature extraction for NLP, simplified. In *International Conference on Language Resources and Evaluation*.

Saqr, M., Poquet, O., & López-Pernas, S. (2022). Networks in education: A travelogue through five decades. *IEEE Access, 10*, 32361–32380. https://doi.org/10.1109/ACCESS.2022.3159674

Scherer, R., & Campos, D. G. (2022). Measuring those who have their minds set: An item-level meta-analysis of the implicit theories of intelligence scale in education. *Educational Research Review, 37*, 100479. https://doi.org/10.1016/j.edurev.2022.100479

Schneider, M., & Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin, 143*(6), 565–600. https://doi.org/10.1037/bul0000098

Schroeder, N. L., & Kucera, A. C. (2022). Refutation text facilitates learning: A meta-analysis of between-subjects experiments. *Educational Psychology Review, 34*(2), 957–987. https://doi.org/10.1007/s10648-021-09656-z

Senden, B., Nilsen, T., & Blömeke, S. (2022). Instructional quality: A review of conceptualizations, measurement approaches, and research findings. In M. Blikstad-Balas, K. Klette, & M. Tengberg (Eds.), *Ways of analyzing teaching quality* (pp. 140–172). Scandinavian University Press. https://doi.org/10.18261/9788215045054-2021-05

Shemilt, I., Khan, N., Park, S., & Thomas, J. (2016). Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews, 5*(1), 140. https://doi.org/10.1186/s13643-016-0315-4

Smith, V., Devane, D., Begley, C. M., & Clarke, M. (2011). Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC Medical Research Methodology, 11*(1), 15. https://doi.org/10.1186/1471-2288-11-15

Tarantino, G., Makopoulou, K., & Neville, R. D. (2022). Inclusion of children with special educational needs and disabilities in physical education: A systematic review and meta-analysis of teachers' attitudes. *Educational Research Review, 36*, 100456. https://doi.org/10.1016/j.edurev.2022.100456

Täschner, J., Dicke, T., Reinhold, S., & Holzberger, D. (2023). "Yes, I can!" A systematic review and meta-analysis of intervention studies promoting teacher self-efficacy. *PsyArXiv*. https://doi.org/10.31234/osf.io/cds45

Taylor, J. A., & Hedges, L. V. (2023). Toward more rapid accumulation of knowledge about what works in physics education: The role of replication, reporting practices, and meta-analysis. In M. F. Taşar & P. R. L. Heron (Eds.), *The international handbook of physics education research: Special topics*. AIP Publishing. https://doi.org/10.1063/9780735425514

Teijema, J., Van de Schoot, R., Ferdinands, G., Lombaers, P., & De Bruin, D. B. (2022). *ASReview Makita: A workflow generator for simulation studies using the command line interface of ASReview LAB* (1.2.0) [Computer software]. https://pypi.org/project/asreview-makita/. https://github.com/asreview/asreview-makita' url: 'https://asreview.ai/

Theobald, M. (2021). Self-regulated learning training programs enhance university students' academic performance, self-regulated learning strategies, and motivation: A meta-analysis. *Contemporary Educational Psychology, 66*, 101976. https://doi.org/10.1016/j.cedpsych.2021.101976

Tsou, A. Y., Treadwell, J. R., Erinoff, E., & Schoelles, K. (2020). Machine learning for screening prioritization in systematic reviews: Comparative performance of Abstrackr and EPPI-Reviewer. *Systematic Reviews, 9*(1), 73. https://doi.org/10.1186/s13643-020-01324-7

Turan, E., & De Smedt, B. (2022). Mathematical language and mathematical abilities in preschool: A systematic literature review. *Educational Research Review, 36*, 100457. https://doi.org/10.1016/j.edurev.2022.100457

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, (2), 125–133. https://doi.org/10.1038/s42256-020-00287-7

van Haastrecht, M., Sarhan, I., Yigit Ozkan, B., Brinkhuis, M., & Spruit, M. (2021). SYMBALS: A systematic review methodology blending active learning and snowballing. *Frontiers in Research Metrics and Analytics, 6*. https://doi.org/10.3389/frma.2021.685591

van Huizen, T., & Plantenga, J. (2018). Do children benefit from universal early childhood education and care? A meta-analysis of evidence from natural experiments. *Economics of Education Review, 66*, 206–222. https://doi.org/10.1016/j.econedurev.2018.08.001

Veletić, J., Rodriguez-Mejia, A. M., & Olsen, R. V. (2023). *A systematic literature review of the Teaching and Learning International Survey (TALIS) research*. Faculty of Educational Sciences. University of Oslo

Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., & Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: A methodological systematic review. *BMC Medical Research Methodology, 19*(1), 132. https://doi.org/10.1186/s12874-019-0782-0

Wagner, S., Schneider, J., & Lachner, A. (2023). *Where to next? Mapping the landscape of research on computerbased feedback on writing*. Faculty of Economic and Social Sciences. University of Tübingen.

Wallace, B. C., Trikalinos, T. A., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC Bioinformatics, 11*(1), 55. https://doi.org/10.1186/1471-2105-11-55

Wang, Y., Tian, J., Yazar, Y., Ones, D. S., & Landers, R. N. (2022). Using natural language processing and machine learning to replace human content coders. *Psychological Methods.* https://doi.org/10.1037/met0000518

World Health Organization. (2019). *The ICD-10 classification of mental and behavioural disorders: Diagnostic criteria for research*. World Health Organization.

Xu, X., Schönrock-Adema, J., Jaarsma, A. D. C., Duvivier, R. J., & Bos, N. A. (2022). A conducive learning environment in international higher education: A systematic review of research on students' perspectives. *Educational Research Review, 37*, 100474. https://doi.org/10.1016/j.edurev.2022.100474

Yu, Z., & Menzies, T. (2019). FAST2: An intelligent assistant for finding relevant papers. *Expert Systems with Applications, 120*, 57–71. https://doi.org/10.1016/j.eswa.2018.11.021

Zhang, Q., Wang, J., & Neitzel, A. (2023). School-based mental health interventions targeting depression or anxiety: A meta-analysis of rigorous randomized controlled trials for school-aged children and adolescents. *Journal of Youth and Adolescence, 52*(1), 195–217. https://doi.org/10.1007/s10964-022-01684-4

Ziernwald, L., Hahnel, C., Reinhold, F., Mitsostergios, G., & Holzberger, D. (2023). *Operationalization and effectiveness of reading strategies in digital reading—A research synthesis*. https://osf.io/2gpzx/

Zinsser, K. M., Silver, H. C., Shenberger, E. R., & Jackson, V. (2022). A systematic review of early childhood exclusionary discipline. *Review of Educational Research, 92*(5), 743–785. https://doi.org/10.3102/00346543211070047

Authors and Affiliations

Diego G. Campos[1] · Tim Fütterer[2] · Thomas Gfrörer[2] · Rosa Lavelle-Hill[2,3] · Kou Murayama[2] · Lars König[4] · Martin Hecht[4] · Steffen Zitzmann[2] · Ronny Scherer[1]

✉ Diego G. Campos
d.g.campos@cemo.uio.no

[1] Centre for Educational Measurement at the University of Oslo (CEMO), Faculty of Educational Sciences, University of Oslo, Oslo, Norway

[2] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

[3] Copenhagen Center for Social Data Science (SODAS) and Department of Psychology, University of Copenhagen, Copenhagen, Denmark

[4] Department of Psychology, Helmut Schmidt University, Hamburg, Germany