

DOI: https://doi.org/10.48009/3_iis_2026_114

AI-assisted versus traditional systematic review approaches: Evidence from accounting education literature

Merve Acar, *Middle Georgia State University*, merve.acar@mga.edu

Tabitha Floyd, *Middle Georgia State University*, tabitha.floyd@mga.edu

Abstract

This study evaluated the degree to which Elicit Pro in Review mode converges with the coverage achieved through traditional expert-driven systematic literature searches and title/abstract screening within the context of accounting education research. A well-established accounting education literature review series served as the benchmark, with Elicit's performance assessed across repeatability, sensitivity, precision, and F1 score metrics, interpreted as measures of convergence with the expert-curated benchmark rather than as absolute indicators of screening accuracy. Results showed substantially lower convergence between Elicit's retrieval and the benchmark across all performance metrics, with precision and sensitivity values consistently in the 1–15% range compared to 98–100% for the benchmark. Notably, overlap between studies retrieved by Elicit and those included in the benchmark was minimal; however, manual screening revealed that a portion of Elicit's non-overlapping retrievals were substantively relevant to accounting education, suggesting that precision estimates relative to the benchmark's five-journal scope may understate Elicit's true retrieval capability. A robustness check using the GALILEO database confirmed that the benchmark's restricted journal scope influences the magnitude of the observed performance gap, though convergence remained limited even under this broader comparison framework. Overall, results indicate that Elicit enhances early-stage exploration and thematic discovery and demonstrates the ability to identify relevant literature beyond the scope of the expert-curated benchmark but does not yet converge sufficiently with traditional review methods for comprehensive literature classification. The two approaches are best understood as complementary within contemporary accounting education research synthesis. Recommendations for future research are provided.

Keywords: systematic reviews, literature reviews, Artificial intelligence tools, AI research tools, AI-assisted research, reliability, accuracy

Introduction

Artificial Intelligence tools are becoming increasingly commonplace in our personal and professional lives. As academics and researchers, AI tools may help streamline various research tasks (van Dijk et al., 2023). Systematic reviews are essential for understanding the evolution, scope, and thematic directions of prior research, yet they are a time-consuming task for researchers (Fabiano et al., 2024). Elicit is an AI-powered research tool specifically designed to aid researchers in the systematic review and data extraction process. Elicit relies on the Semantic Scholar database of more than 126 million academic papers to quickly generate a summary of a posed research question and find prior research related to that question (Lau & Golder, 2025). Artificial intelligence tools aim to simplify and accelerate the systematic review process without

compromising accuracy; however, empirical evaluations of their performance remain limited (Lau & Golder, 2025; Bernard et al., 2025).

This study evaluated the screening performance of the AI-assisted literature review tool Elicit Pro subscription in Review mode relative to a traditional systematic review approach. To assess this, the research compared the studies identified and screened by Elicit with those included in Apostolou et al.'s well-established accounting education literature review series, which provides one of the most comprehensive classifications of accounting education scholarship from 1985 to 2024 (Churyk et al., 2025, 2024; Apostolou et al., 2023, 2022, 2021, 2020, 2019, 2018, 2017, 2016, 2015, 2013, 2010, 2001; Watson et al., 2007, 2003; Rebele et al., 1998a, 1998b; Rebele et al., 1991; Rebele and Tiller, 1986). The research calculated repeatability, accuracy, sensitivity, and precision scores to assess Elicit's ability to replicate coverage and identify relevant literature. This study will answer the following research question:

RQ1: Does systematic review screening with Elicit Pro add value to the process compared to traditional systematic review screenings?

Literature Review

Systematic Reviews

A systematic review is a structured overview of available research on a specific topic or research question. The review should be complete and thorough and thus help researchers identify gaps in the literature for future research (Fabiano et al., 2024). Systematic reviews are critical to advancing scholarship, yet they are widely recognized as a time-consuming and demanding process. There are many different stages of a systematic review – starting with developing the research question, identifying and selecting relevant studies through systematic review of databases, reviewing the included studies for quality checks, summarizing the evidence from the studies, and interpreting the findings (Fabiano et al., 2024). AI tools such as Elicit are offered as a potential solution to aid in these various stages of the review process (Bernard et al., 2025; Lau & Golder, 2025; Fabiano et al., 2024; van Dijk et al., 2023). Prior studies of the use of AI tools within the systematic review screening process have shown that the use of AI tools makes the process more efficient by reducing the number of articles in the screening process, and thus, saving a great deal of time (van Dijk et al., 2023). Although there have been recent studies related to the effectiveness of Elicit Pro related to health sciences systematic review processes, research on the effectiveness of Elicit and other AI tools remains limited (Bernard et al., 2025; Lau & Golder, 2025).

Apostolou's Benchmark Review

The benchmark for comparison was the well-established accounting education literature review series conducted by Apostolou et al. (2023), which provides a comprehensive and curated synthesis of accounting education research over multiple decades. The Apostolou et al. series was selected as the benchmark because it represents one of the most methodologically rigorous syntheses of accounting education research available in literature. The researchers employed structured and transparent procedures for identifying, classifying, and summarizing research, making the series an appropriate reference standard for evaluating AI-assisted screening tools. This review series systematically synthesizes accounting education research across five major themes including curriculum and instruction, instruction by content area, educational technology, students, and faculty. This series includes articles published in five accounting education journals: *Journal of Accounting Education*, *Accounting Education*, *Advances in Accounting Education: Teaching and Curriculum Innovations*, *Issues in Accounting Education*, and *The Accounting Educator's Journal* (Churyk et al., 2025, 2024; Apostolou et al., 2023, 2022, 2021, 2020, 2019, 2018, 2017, 2016, 2015,

2013, 2010, 2001; Watson et al., 2007, 2003; Rebele et al., 1998a, 1998b; Rebele et al., 1991; Rebele and Tiller, 1986).

Additional justification for selecting Apostolou et al.'s review series as the benchmark derives from the significant scholarly influence of its authors within the accounting education field. Bibliometric evidence identifies Apostolou, Hassell, and Rebele among the most impactful contributors to accounting education research based on publication output and citation counts. Hassell ranks highest across statistical indicators, with 18 publications and 842 citations between 1991 and 2023, demonstrating both sustained productivity and scholarly influence. Rebele and Apostolou follow closely, each contributing 16 publications to the field (Hu et al., 2025). The high citation impact and long-standing contributions of these authors indicate that the studies included in their review series reflect widely recognized and influential scholarship. Consequently, using this review series as the benchmark strengthens the evaluation's validity by anchoring the comparison in a curated body of literature developed by leading experts.

AI Tools for Academic Literature Reviews

Recent publications summarize the available AI tools for researchers and how these tools can be incorporated into the systematic review process to improve both efficiency and quality (Fabiano et al., 2024). There is growing interest in exploring how these AI tools can help researchers (Bernard et al., 2025). These tools have the potential to assist in the varied tasks of the systematic review process (Fabiano et al., 2024). Both Elicit and Consensus allow researchers to ask research questions and then generate screening criteria and provide a response to the prompt, including summaries of prior research on the topic (Lau & Golder, 2025; Fabiano et al., 2024). Research Rabbit can suggest publications based on saved files or from a title which helps with discovery of new articles. Other AI tools – such as PaperDigest and ChatGPT - can summarize an article and extract data from the text (Fabiano et al., 2024). These tools are continually improving (Lau & Golder, 2025).

Elicit uses the Semantic Scholar database to identify relevant publications, providing access to more than 126 million academic papers and relies on semantic similarity to identify papers relevant to a posed research question. Elicit then generates a summary to the research question through an analysis of each identified paper's abstract (Bernard et al., 2025). This report provides a comprehensive summary of the topic. Elicit allows users to refine the search by applying filters for specific keywords, article types, and publication years. Additionally, Elicit displays results by showing key information such as population and methodology in a table format on the eight most relevant studies (Bernard et al., 2025). Elicit gives a "show more" feature to retrieve additional papers. Although Elicit enables a highly automated screening process, the platform allows users to manually review and override inclusion decisions, when necessary (Byun, 2025; Lau & Golder, 2025), thereby preserving researcher judgment in the final study selection. Additionally, Elicit provides a spreadsheet with columns for each inclusion criteria and whether the study meets that criteria increasing transparency in the process (Lau & Golder, 2025). Elicit uses the Semantic Scholar database, so this limits the possible articles found if studies are not referenced in that database (Bernard et al., 2025). Elicit provides three tiers of accounts and three different modes. The Review mode, designed for systematic reviews, is only available in the highest level (Pro) subscription-based account (Lau & Golder, 2025).

Using AI tools in systematic reviews can pose certain risks with accuracy, reliability, and creditability. With the increased use of AI in research, research integrity becomes even more vital to ensure reliability of the results (Lau & Golder, 2025). The risk that the AI tool could generate false or misleading evidence necessitates that human expertise, verification and judgement remain a necessary component for quality research. Prior studies conclude that AI tools should be viewed as a supplement to the expert and not a replacement for the expert (Lau & Golder, 2025; Bernard et al., 2025; Fabiano et al., 2024). Notably, prior research shows that Elicit Pro finds relevant articles not identified by traditional systematic review

procedures thus helping to increase the completeness of the review (Lau & Golder, 2025; Bernard et al., 2025). The expert works with AI tools to minimize risks, producing both quality research and an efficient process (Fabiano et al., 2024). Researchers using AI tools should always report the AI tools used, the procedures used to verify the AI's accuracy, and the settings of the AI tools to ensure replicability. The burden of correcting errors made by the AI tool is on the researchers and cannot be deferred to AI (Fabiano et al., 2024).

Performance Metrics

In systematic reviews, screening performance can be evaluated using information retrieval metrics that quantify how effectively a method identifies relevant studies. In this paper, the abstract screening performance of Elicit versus benchmark review series was assessed based on two criteria: repeatability and accuracy (Kitchenham et al., 2011; Bernard et al., 2025; Lau & Golder, 2025). Repeatability was defined as Elicit's ability to produce consistent results under identical conditions at different points in time (Bernard et al., 2025).

Precision, sensitivity (recall), and F1 score are widely used measures to assess the accuracy of study identification relative to a benchmark or reference dataset. Precision represents the proportion of studies identified as relevant that are truly relevant, reflecting the extent of Type I error (false positives) in the screening process. It reflects how efficiently the screening process excludes irrelevant studies and was calculated as:

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

Where TP stands for "True Positives" that are correctly identified as relevant studies by the screening tool, and FP stands for "False Positives" for studies identified as relevant by the tool but not included in the benchmark review. Precision, therefore, indicates the proportion of retrieved studies that truly meet the inclusion criteria (Boeker et al., 2013). A high precision score indicates that most of the studies identified by the tool are relevant, meaning the screening method produces relatively few false positives. This reduces the amount of manual effort required to review irrelevant studies and improves the efficiency of the review process. On the other hand, a low precision score indicates that many irrelevant studies have been incorrectly classified as relevant. In systematic reviews, low precision increases the workload of reviewers because more studies must be manually screened and excluded.

Sensitivity, also referred to as Recall, represents the proportion of truly relevant studies that were correctly identified, reflecting Type II error (false negatives), and was calculated as follows:

$$\text{Sensitivity (Recall)} = \frac{TP}{(TP+FN)}$$

In this formula, TP stands for "True Positives" that are correctly identified by the screening method, and FN stands for "False Negatives" that are relevant studies that were not identified. A high sensitivity value indicates that the screening approach captures most of the relevant literature, minimizing the number of missed studies. Sensitivity is particularly critical in systematic reviews because failing to identify relevant studies may introduce selection bias and threaten the validity of conclusions. Evidence synthesis relies on comprehensive coverage of the literature, and missing key studies may distort estimates, thematic interpretations, or theoretical conclusions (Chandler et al., 2019). For this reason, systematic review guidelines typically prioritize high sensitivity during the search stage, even if this leads to the retrieval of some irrelevant studies that must later be excluded during screening (Lefebvre et al., 2019). Moreover, sensitivity is directly related to Type II error, which occurs when a relevant study is incorrectly excluded.

A low sensitivity value indicates a high number of false negatives, meaning that important studies may have been overlooked (Grandini et al., 2020). Because systematic reviews aim to minimize the risk of incomplete evidence synthesis, reducing false negatives is often considered more important than minimizing false positives during the initial screening stage (Kitchenham et al., 2009).

While precision and sensitivity scores provide insight into different types of classification errors, they should be interpreted together because optimizing one may reduce the other. For example, a method that classifies all studies as relevant would achieve perfect sensitivity but very low precision. To provide a single performance indicator, precision and recall are commonly combined into the F1 score, which is calculated as the harmonic mean of the two measures. In systematic reviews, the F1 score is particularly useful because it evaluates the balance between identifying relevant studies (high sensitivity) and avoiding inclusion of irrelevant studies (high precision). This balance is important when assessing AI-assisted literature screening tools, where both completeness and efficiency are desired.

The F1 score was calculated as the harmonic mean of precision and sensitivity:

$$F1\ score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

The harmonic mean was used because it penalizes extreme imbalances between precision and recall, ensuring that the F1 score remains low when either measure is low. For example, when precision is 0.9 and recall is 0.1, the F1 score equals 0.18, whereas the arithmetic mean would yield a misleading value of 0.5. Therefore, the F1 score provided a balanced metric for evaluating screening performance (Hossin and Sulaiman, 2015; Grandini et al., 2020; Owen et al. 2026). Systematic review screening methods often involve a trade-off between sensitivity and precision. A highly sensitive search strategy may retrieve most relevant studies but also produce many irrelevant results, increasing screening workload. Conversely, a highly precise strategy may reduce irrelevant results but risk missing relevant studies. The F1 score provides a single summary measure that reflects the balance between these competing objectives. This makes it particularly useful when comparing different literature search approaches, including AI-assisted tools and traditional database searches (Van Rijsbergen, 1979). Because the F1 score incorporates both precision and sensitivity, it reflects both Type I error (captured by precision) and Type II error (captured by sensitivity) (Christen et al., 2023; Powers, 2020).

Table 1 summarizes performance metrics used to measure accuracy of Elicit screening results to traditional systematic review.

Table 1. Performance Metrics for Evaluating Systematic Review Screening: AI (Elicit) versus Traditional Approach

Metric	Formula	Definition	Interpretation in Systematic Review
Repeatability	Consistency of retrieved study sets across repeated searches	Ability of the search tool to produce stable results under identical search conditions at different times.	High repeatability indicates reliability and stability of the search process.
Sensitivity (Recall)	$TP / (TP + FN)$	Proportion of truly relevant studies that were correctly identified by the screening method.	Indicates completeness of literature coverage. High sensitivity means few relevant studies were missed (low false negatives).
Precision	$TP / (TP + FP)$	Proportion of retrieved studies that are truly relevant.	Indicates efficiency of screening. High precision means few irrelevant studies were included (low false positives).
F1 Score	$2 \times (\text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity})$	Harmonic mean of precision and sensitivity, providing a single measure of overall screening performance.	Indicates balance between completeness (recall) and relevance (precision). High F1 score suggests strong overall screening accuracy.

Methodology

Since Elicit relies on semantic search rather than conventional keyword-based search strings typically used in systematic reviews, the first step involves formulating a structured research question. Consistent with the thematic framework used in the benchmark review series by Apostolou et al., the research question was developed to reflect key dimensions of accounting education research following a Population Intervention Context Outcomes and Study Design (PICOS) informed logic adapted for educational research (Stone, 2002; Paynter, 2009). The PICO framework is widely used to structure research questions in quantitative systematic reviews. It helps clearly define the key components of a review question and ensures conceptual precision during study identification and screening. The framework is strongly endorsed by the Cochrane Collaboration, a leading authority on evidence-based research methods, and is recommended in the *Cochrane Handbook for Systematic Reviews of Interventions* as a model for developing well-structured review questions (Higgins and Green, 2006; Eriksen and Frandsen, 2018). In line with the PICO framework, population can be defined as “accounting education literature,” intervention as “instructional approaches and educational innovations,” context as “higher education accounting programs,” and Outcomes as “evolution of research themes.” In line with the PICO approach, the following query was entered into Elicit: “How has accounting education research evolved over time in areas such as curriculum design, instructional methods, educational technology, student learning, and faculty roles?” This structured question ensures conceptual alignment between the Elicit search process and the thematic categories used in the benchmark review, enabling meaningful comparison of screening performance between AI-assisted and traditional literature review approaches.

After formulating the research question and entering the query into Elicit, filters were applied to refine the search results based on publication years (1997–2024) and relevance to peer-reviewed academic journal

articles (RCT-Randomized Clinical Trials). The benchmark review covered the period from 1997 to 2024. To enhance coverage, the Elicit search was conducted using year-based intervals: 1997–2005, 2006–2012, 2013–2019, 2020, 2021, 2022, 2023, and 2024. For earlier years, separate year-by-year searches were not performed because the overlap between studies identified by Elicit and those included in the benchmark review was very limited prior to 2018. Due to this low level of common studies, results for the 1997–2018 period were grouped into three broader subperiods (1997–2005, 2006–2012, and 2013–2018) to improve interpretability of the comparison and reduce duplication.

Based on the research query, Elicit identified the 500 most relevant studies using semantic similarity of titles and abstracts. At the time of the initial analysis, Elicit retrieved and screened the 500 most relevant studies based on semantic similarity of titles and abstracts. However, during the final stages of this study, Elicit updated its system to allow screening of up to 1,000 studies, thereby expanding the pool of potentially eligible literature identified through the semantic search process. Accordingly, only the last trial of analysis used the expanded 1,000-study retrieval set, while earlier trials were limited to 500 studies. Because some of the automatically generated criteria did not fully align with the inclusion criteria applied in the benchmark review, the authors refined and adjusted the screening criteria within Elicit to better match the conceptual scope of the Apostolou et al. review series. After screening was completed, all 500 candidate studies were exported into Microsoft Excel for further comparison with the studies included in the benchmark review. Manual screening ensured conceptual alignment between Elicit results and the benchmark study's thematic boundaries. Inclusion criteria were aligned with the benchmark review study and included:

1. Studies focusing on accounting education research
2. Articles addressing curriculum, pedagogy, instructional innovation, technology use, assessment, or faculty development
3. Peer-reviewed journal publications
4. Studies published within the time periods covered by the benchmark review series
5. Empirical or conceptual contributions relevant to accounting education scholarship

Exclusion criteria included:

- studies unrelated to accounting education
- editorials, commentary, or non-scholarly publications
- duplicate records
- studies outside the thematic scope of the benchmark review

Based on the Elicit screening, manual screening and the comparison between Elicit results and the benchmark review, the authors categorized studies into four groups:

1. Studies included in the benchmark review but not identified by Elicit
2. Studies identified by Elicit but excluded after applying screening criteria
3. Studies identified and included by both Elicit and the benchmark review
4. Studies identified by Elicit but not included in the benchmark review

The researchers evaluated whether these represented potentially relevant additional contributions based on the original inclusion criteria. This step enabled assessment of whether AI-assisted screening could identify relevant studies not captured in the traditional review process.

In this paper, the abstract screening performance of Elicit versus benchmark review series was assessed based on two criteria (Kitchenham et al., 2011; Bernard et al., 2025; Lau & Golder, 2025): repeatability and accuracy. In this study, repeatability was defined as Elicit's ability to produce consistent results under

identical conditions at different points in time. To evaluate repeatability, the search procedure described above was replicated three times at different times (varying by hour and day) using the same research query. The results obtained from these trials were then compared to assess the consistency of the retrieved studies.

In addition to the reliability measures, accuracy was measured and analyzed using relevance (sensitivity), precision, and F1 score metrics. Precision and sensitivity metrics were calculated using the number of common papers (true positives) and the number of irrelevant papers retrieved by Elicit (false positives), as well as relevant benchmark studies not identified by Elicit (false negatives). These measures enabled assessment of Elicit's ability to accurately retrieve relevant literature while minimizing inclusion of non-relevant studies. Because the Apostolou et al. benchmark is not an exhaustive census of accounting education literature but rather a scope-limited expert synthesis, precision and sensitivity in this study are interpreted as measures of convergence between Elicit's retrieval and the benchmark rather than as absolute indicators of accuracy. Studies retrieved by Elicit but absent from the benchmark are classified as non-overlapping retrievals rather than strict false positives in the traditional sense, as some may represent relevant literature outside the benchmark's five-journal scope.

Robustness Checks

As an additional robustness test, the authors compared the sensitivity and precision metrics obtained from Elicit and the benchmark review series with results derived from database searches conducted through the library system of a public state university. The purpose of this analysis was to evaluate whether Elicit's performance is influenced by the scope of journals included in the benchmark review and to assess whether broader database coverage affects the degree of overlap between AI-assisted and traditional literature search results.

The benchmark review series focus primarily on five core journals in accounting education (Churyk et al., 2025, 2024; Apostolou et al., 2023, 2022, 2021, 2020, 2019, 2018, 2017, 2016, 2015, 2013, 2010, 2001; Watson et al., 2007, 2003; Rebele et al., 1998a, 1998b; Rebele et al., 1991; Rebele and Tiller, 1986). While this targeted scope ensures consistency across review periods, it may also limit the breadth of studies included in the benchmark dataset. In contrast, the university library databases provide access to a wider range of journals and publication outlets. When comparing Elicit results with studies retrieved through these broader database searches, the number of overlapping studies increased relative to the comparison with the benchmark review alone. Precision for the benchmark review is consistently 100%, as all studies included in the benchmark dataset are directly focused on accounting education and therefore fully meet the relevance criteria. Consequently, sensitivity and precision values for Elicit improved under this expanded comparison framework.

Results

Results on Repeatability

As described above, repeatability was defined as the ability of the AI screening tool to produce consistent results under identical conditions at different points in time. To evaluate repeatability, the search procedure described above was conducted three times using the same research query, with each search performed at different times (varying by hour and day). The searches were carried out between December 2025 and February 2026. The results of these repeated trials are presented in Table 2. The results indicate that the screening outcomes generated by Elicit were not fully consistent across repeated trials conducted under identical search conditions. Although the same research question and filters were used in each trial, the number of retrieved studies and the composition of the screened results varied across searches performed at different times between December 2025 and February 2026.

Table 2. Number of studies retrieved across repeated Elicit search trials

Trial	Trial 1		Trial 2		Trial 3	
Years	<i>Search Date / Time</i>	<i>Number of papers</i>	<i>Search Date / Time</i>	<i>Number of papers</i>	<i>Search Date / Time</i>	<i>Number of papers</i>
1997- 2005	01/16/2026 2:09 pm	127	02/18/2026 6:35 pm	103	02/13/2026 at 6:19 pm	65
2006- 2014	01/16/2026 2:52 pm	136	02/18/2026 5:10 pm	98	02/13/2026 at 5:24 pm	73
2015- 2018	01/16/2026 2:32 pm	106	02/18/2026 12:11 pm	85	02/13/2026 at 5:33 pm	68
2019	12/12/2025 2:53 pm	94	12/13/2025 1:56 pm	70	12/14/2025 4:24 pm	100
2020	12/12/2025 9:49 am	83	12/13/2025 2:05 pm	72	12/14/2025 5:44 pm	109
2021	12/12/2025 10:09 am	86	12/13/2025 3:26 pm	77	12/14/2025 5:50 pm	112
2022	12/12/2026 11:44 am	87	12/13/2026 3:45 pm	75	12/14/2025 6:04 pm	106
2023	12/12/2026 11:36 am	55	12/13/2026 2:11 pm	68	12/27/2025 3:13 pm	122
2024	5/11/2026 10:51 pm	34	5/12/2026 9:00 pm	51	5/12/2026 9:40 pm	103
Total		439		413		652

Results on Accuracy

Table 3 Panel A summarizes the agreement between Elicit and the classical screening method. Based on manual screening and comparison with the benchmark review, the overlap between studies identified by Elicit and those included in the benchmark was limited, with relatively few common papers identified.

Table 3, Panel B presents the precision, sensitivity (recall), and F1 scores calculated for each trial. Due to the very small number of overlapping studies between Elicit and the benchmark review prior to 2019, repeatability analysis for the 1997–2018 period was based on a single trial. The limited overlap observed in earlier years also supports the finding that Elicit’s screening performance improves for more recent publications. To maintain brevity, only years showing at least one common study between Elicit and the benchmark review series are reported. Across all trials, the performance of Elicit was consistently lower than that of the benchmark traditional review.

Table 3 Panel A. Overlap and differences between studies identified by Elicit and the traditional review

Trial 1			
Years	Elicit Inclusion (met screening criteria)	Elicit truly relevant (accounting education centered)	# of Common papers
1997-2005	127	14	0
2006-2014	177	51	0
2015-2019	204	123	6
2020	93	31	1
2021	91	46	2
2022	93	33	1
2023	62	21	1
2024	34	22	2
Trial 2			
Years	Elicit Inclusion (met screening criteria)	Elicit truly relevant (accounting education centered)	# of Common papers
1997-2005	127	14	0
2006-2014	177	51	0
2015-2019	177	106	6
2020	78	23	1
2021	79	42	1
2022	78	30	1
2023	69	30	1
2024	51	22	2
Trial 3			
Years	Elicit Inclusion (met screening criteria)	Elicit truly relevant (accounting education centered)	# of Common papers
1997-2005	127	14	0
2006-2014	177	51	0
2015-2018	206	101	6
2020	109	17	0
2021	112	41	2
2022	106	18	1
2023	122	43	1
2024	103	46	3

Table 3 Panel B. Accuracy metrics (Sensitivity, Precision, and F1 Score) comparing Elicit screening results with the benchmark review

Trial 1 Years	Sensitivity- Elicit	Sensitivity- Benchmark	Precision- Elicit	Precision- Benchmark	F1 Score- Elicit	F1 Score- Benchmark
2015-2018	1.83%	100%	15.15%	100%	3.27%	100%
2019	3.57%	98.21%	3.70%	100.00%	3.64%	99.10%
2020	1.49%	100.00%	1.69%	100.00%	1.59%	100.00%
2021	3.61%	98.78%	6.38%	100.00%	4.62%	99.39%
2022	3.17%	98.41%	3.23%	100.00%	3.20%	99.20%
2023	1.32%	100.00%	2.38%	100.00%	1.69%	100.00%
2024	2.63%	100.00%	14.29%	100.00%	1.69%	100.00%
Trial 2 Years	Sensitivity- Elicit	Sensitivity- Benchmark	Precision- Elicit	Precision- Benchmark	F1 Score- Elicit	F1 Score- Benchmark
2015-2018	1.83%	100%	15.15%	100%	4.44%	100%
2019	3.57%	98.21%	6.67%	100.00%	4.65%	99.10%
2020	1.49%	100.00%	1.79%	100.00%	1.63%	100.00%
2021	2.44%	98.78%	5.26%	100.00%	3.33%	99.39%
2022	3.17%	98.41%	4.00%	100.00%	3.54%	99.20%
2023	1.32%	100.00%	2.50%	100.00%	1.72%	100.00%
2024	2.63%	100.00%	6.45%	100.00%	3.74%	100.00%
Trial 3 Years	Sensitivity- Elicit	Sensitivity- Benchmark	Precision- Elicit	Precision- Benchmark	F1 Score- Elicit	F1 Score- Benchmark
2015-2018	1.83%	100%	15.15%	100%	3.27%	100%
2019	3.57%	98.21%	2.53%	100.00%	2.96%	99.10%
2020	0.00%	100.00%	0.00%	100.00%	0.00%	100.00%
2021	3.66%	98.78%	4.11%	100.00%	3.87%	99.39%
2022	3.17%	98.41%	2.22%	100.00%	2.61%	99.20%
2023	1.32%	100.00%	1.25%	100.00%	1.28%	100.00%
2024	3.95%	100.00%	5.0%	100.00%	4.41%	100.00%

Results on Robustness checks

To provide a more conservative estimate of false positives, we recalculated precision using the subset of Elicit-retrieved studies judged as substantively relevant to accounting education during manual screening, rather than restricting true positives solely to benchmark-included studies. This adjusted precision estimate better reflects Elicit's actual relevance filtering capability independent of the benchmark's journal-scope constraints. Table 4 summarizes the screening performance of Elicit and benchmark review series using the alternative database search for the robustness check. To directly address the potential bias introduced by the benchmark's limited journal set, we conducted a robustness check comparing Elicit's results against a broader university library database search encompassing a wider range of journals. Results show that Elicit's precision and sensitivity improved under this expanded comparison, and the benchmark's own sensitivity decreased, confirming that the five-journal scope does influence the magnitude of the performance gap. Importantly, however, even under the broader comparison framework, Elicit's sensitivity remained well

below 11% across all periods, and precision remained low. To maintain brevity, results are reported only for years in which Elicit produced at least one study in common with the library database search.

Table 4. Comparison of screening performance across Elicit, benchmark review, and database search

Years	Sensitivity- Elicit	Sensitivity- Benchmark Review	Precision- Elicit	Precision- Benchmark Review	F1 Elicit	Score- Benchmark Review
2015- 2018	10.20%	75.51%	15.15%	100.00%	12.20%	86.05%
2019	6.67%	53.33%	3.70%	100.00%	4.76%	69.57%
2020	8.33%	61.11%	4.92%	100.00%	6.19%	75.86%
2021	3.13%	50.00%	2.38%	100.00%	2.70%	66.67%

With respect to the research question, the findings of this study provide consistent evidence that Elicit may show limited convergence with the benchmark or retrieve a substantially different set of studies than the benchmark. Low precision scores reflect limited convergence between Elicit's retrievals and the benchmark's journal-scoped inclusion criteria, with a substantial proportion of Elicit-retrieved studies falling outside the benchmark's five-journal scope. Collectively, the results suggest that, although Elicit may offer efficiency and accessibility advantages, its effectiveness in identifying and retrieving relevant literature remains substantially weaker than that of the traditional systematic review method within the context of this study.

Discussion

The performance metrics reported in this study should be interpreted as indicators of the degree of overlap between Elicit's retrieval and the Apostolou et al. benchmark rather than as definitive measures of Elicit's absolute screening accuracy. This distinction is particularly relevant when interpreting precision, as studies retrieved by Elicit but absent from the benchmark may include relevant accounting education literature published outside the benchmark's five-journal scope. That being said, the findings on repeatability suggest that Elicit's semantic search process may produce slightly different results over time, potentially due to ongoing updates in the underlying database (Semantic Scholar), algorithm refinements, or changes in indexing and relevance scoring. While the overall thematic relevance of retrieved studies remained broadly consistent, variation in the specific set of identified studies indicates limited repeatability of screening results across trials. This variability should be considered when interpreting AI-assisted literature review outcomes and highlights the importance of documenting search dates and procedures to enhance transparency and reproducibility.

Both precision and sensitivity values indicate that Elicit identified only a limited proportion of the studies included in the benchmark review, while also retrieving a number of studies that did not meet the benchmark inclusion criteria. As a result, the combined F1 scores remained relatively low across trials, reflecting an imbalance between correctly identified studies and irrelevant or missed studies.

The relatively low sensitivity values suggest that Elicit did not capture a substantial share of studies included in the benchmark review, reflecting limited convergence between Elicit's retrieval and the benchmark's expert-curated, journal-scoped selection. At the same time, moderate to low precision values indicate that a substantial proportion of studies retrieved by Elicit were not included in the benchmark review, representing non-overlapping retrievals that fall outside the benchmark's five-journal scope rather than

strictly irrelevant results. It should be noted that manual screening revealed that some of these non-overlapping retrievals were substantively relevant to accounting education, suggesting that precision estimates in this study may understate Elicit's true retrieval capability when evaluated against a broader reference set. Together, these results indicate that Elicit's retrieval showed substantially lower convergence with the expert-curated benchmark than the traditional review approach, reflecting the fundamental differences in scope and methodology between semantic AI-assisted search and expert-driven journal-specific literature classification.

However, the analysis also revealed instances in which Elicit identified studies that were not included in the traditional review but appeared relevant based on the original inclusion criteria. These findings suggest that AI-assisted semantic search may have the potential to identify additional relevant literature that may have been overlooked in manual review processes. This finding is consistent with the convergence interpretation of our metrics: studies identified by Elicit but absent from the benchmark are not uniformly irrelevant but rather reflect the different retrieval scope of the two approaches. Therefore, although overall performance metrics indicate lower alignment with the benchmark review, Elicit may still provide value as a complementary tool by expanding coverage and identifying potentially relevant studies beyond those captured through traditional search methods.

Overall, the results indicate that while Elicit offers a transparent and efficient screening process, its retrieval showed substantially lower convergence with the expert-curated benchmark than the traditional literature review approach, reflecting fundamental differences in scope and methodology rather than solely differences in screening quality. At the same time, Elicit's ability to identify substantively relevant studies beyond the benchmark's five-journal scope suggests that AI-assisted tools may be particularly useful as a supplementary approach to support and enhance traditional systematic review workflows, potentially expanding the breadth of literature coverage in ways that complement expert-driven methods.

Despite this improvement, Elicit's convergence with the traditional database search approach remained limited. One explanation is that broader journal coverage increases the likelihood that Elicit retrieves studies indexed across multiple sources, thereby increasing the number of overlapping studies. In contrast, the benchmark review's restriction to five core journals reduces the probability of overlap with Elicit's cross-disciplinary retrieval, even when the studies Elicit identifies are substantively relevant to accounting education. This distinction underscores the importance of interpreting our precision and sensitivity metrics as measures of convergence with a scope-limited benchmark rather than as absolute indicators of Elicit's screening accuracy.

Additionally, alternative wording of the research query was tested to assess whether Elicit's semantic search results are sensitive to phrasing differences (Bernard et al., 2025). Because Elicit relies on natural language processing rather than predefined keyword strings, although some variation in results was observed, the general pattern remained consistent.

Overall, the robustness check supports the main findings of the study. While Elicit's sensitivity and precision remain below those of traditional database searches, its performance improves when evaluated against a broader journal set rather than a limited group of benchmark journals. Although there is no universally fixed cutoff point for sensitivity in systematic reviews in the social sciences, methodological literature on systematic review performance generally emphasizes that high sensitivity is preferred, often ≥ 0.80 (80%), and ideally ≥ 0.90 (90%), depending on the review purpose (Lau & Golder, 2025; Higgins et al., 2006; Kitchenham et al., 2009). However, both the main results and the robustness check indicate very limited sensitivity performance (below 10%). Such low sensitivity implies that a substantial proportion of relevant studies were not identified by Elicit. Therefore, in its current form, Elicit should not be relied upon

as a stand-alone tool for systematic review searches. Instead, it may be more appropriately used as a complementary tool alongside traditional database search methods to support literature identification and screening. To summarize, these findings suggest that AI-assisted screening tools may provide additional value when used alongside comprehensive database searches and when research queries are carefully formulated.

Limitations

A key limitation of this study concerns the use of the Apostolou et al. review series as the reference standard for evaluating Elicit's screening performance. While this benchmark represents one of the most methodologically rigorous and influential syntheses in accounting education research, it is not an exhaustive census of all relevant accounting education literature. Its scope is intentionally restricted to five core accounting education journals, meaning that studies published in other outlets (even those substantively relevant to accounting education) fall outside the benchmark by design. As a result, precision and sensitivity metrics in this study are better interpreted as measures of convergence between Elicit's retrieval and the expert-curated benchmark rather than as absolute measures of accuracy. In particular, Elicit-retrieved studies that do not appear in the benchmark are classified as false positives in our metric framework; however, manual screening revealed that a subset of these studies were substantively relevant to accounting education, suggesting that our precision estimates for Elicit may be understated. To partially address this limitation, two corrective steps were taken. First, a robustness check was conducted comparing Elicit's performance against a broader university library database search encompassing a wider journal set, which confirmed that the benchmark's journal-scope constraint influences the magnitude of the performance gap, though Elicit's sensitivity remained well below acceptable thresholds even under this expanded comparison. Second, manual screening results (Table 3, Panel A) allow readers to distinguish between studies excluded solely due to benchmark scope and those judged as substantively irrelevant, providing a more nuanced basis for interpreting precision estimates. Future research would benefit from using an independent, comprehensively constructed reference set spanning a broader range of journals to provide a less scope-constrained evaluation of AI-assisted screening tools in accounting education.

Conclusion

In conclusion, Elicit Pro has the potential to aid researchers in the screening process of systematic reviews. At the time of this evaluation, this study showed that Elicit was not as accurate as an expert led systematic review process. Due to its low sensitivity, it should not be the sole source of the review materials. Our results, consistent with prior research, show Elicit should be used as enhancement tool to traditional methods, instead of replacing traditional methods (Lau & Golder, 2025; Fabiano et al., 2024). Importantly, our findings, consistent with prior research, show that Elicit was able to find additional relevant articles that were not included in the benchmark study. This ability reinforces Elicit's value in ensuring the systematic review is comprehensive (Lau & Golder, 2025; Bernard et al., 2025).

Improvements to AI tools are happening rapidly. Further research should continue as new advancements in this technology occur. Further research should include examining whether different wording of the research question impacts Elicit's performance. The researchers encourage future studies evaluating the performance of other AI tools in the systematic review process and within different academic disciplines. Additionally, as with any use of AI in our professional lives, the researchers encourage transparent reporting of the use of AI to ensure replicability of the research.

References

- Apostolou, B., Churyk, N. T., Hassell, J. M., & Matuszewski, L. (2023). Accounting education literature review (2022). *Journal of accounting education*, 63, 100831.
- Apostolou, B., Dorminey, J. W., & Hassell, J. M. (2022). Accounting education literature review (2021). *Journal of Accounting Education*, 59, 100781.
- Apostolou, B., Dorminey, J. W., & Hassell, J. M. (2021). Accounting education literature review (2020). *Journal of Accounting Education*, 55, 100725.
- Apostolou, B., Dorminey, J. W., & Hassell, J. M. (2020). Accounting education literature review (2019). *Journal of Accounting Education*, 51, 100670.
- Apostolou, B., Dorminey, J. W., Hassell, J. M., & Hickey, A. (2019). Accounting education literature review (2018). *Journal of Accounting Education*, 47, 1-27.
- Apostolou, B., Dorminey, J. W., Hassell, J. M., & Rebele, J. E. (2018). Accounting education literature review (2017). *Journal of accounting education*, 43, 1-23.
- Apostolou, B., Dorminey, J. W., Hassell, J. M., & Rebele, J. E. (2017). Accounting education literature review (2016). *Journal of Accounting Education*, 39, 1-31.
- Apostolou, B., Dorminey, J. W., Hassell, J. M., & Rebele, J. E. (2016). Accounting education literature review (2015). *Journal of Accounting Education*, 35, 20-55.
- Apostolou, B., Dorminey, J. W., Hassell, J. M., & Rebele, J. E. (2015). Accounting education literature review (2013–2014). *Journal of Accounting Education*, 33(2), 69-127.
- Apostolou, B., Dorminey, J. W., Hassell, J. M., & Watson, S. F. (2013). Accounting education literature review (2010–2012). *Journal of Accounting Education*, 31(2), 107-161.
- Apostolou, B., Hassell, J. M., Rebele, J. E., & Watson, S. F. (2010). Accounting education literature review (2006–2009). *Journal of Accounting Education*, 28(3-4), 145-197.
- Apostolou, B., Watson, S. F., Hassell, J. M., & Webber, S. A. (2001). Accounting education literature review (1997–1999). *Journal of Accounting Education*, 19(1), 1-61.
- Bernard, N., Sagawa Jr, Y., Bier, N., Lihoreau, T., Pazart, L., & Tannou, T. (2025). Using artificial intelligence for systematic review: the example of elicitor. *BMC medical research methodology*, 25(1), 75.

- Boeker, M., Vach, W., & Motschall, E. (2013). Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. *BMC medical research methodology*, 13(1), 131.
- Byun, J. (2025, February 20). *Introducing Elicit Systematic Review*. The Elicit Blog. <https://blog.elicit.com/systematic-review/>.
- Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. J. H. W. (2019). Cochrane handbook for systematic reviews of interventions. *Hoboken: Wiley*, 4(1002), 14651858.
- Christen, P., Hand, D. J., & Kirielle, N. (2023). A review of the F-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3), 1-24.
- Churyk, N. T., Eaton, T. V., & Matuszewski, L. J. (2024). Accounting education literature review (2023). *Journal of Accounting Education*, 67, 100901.
- Churyk, N. T., Eaton, T. V., & Matuszewski, L. J. (2025). Accounting education literature review (2024). *Journal of Accounting Education*, 71, 100969.
- Eriksen, M. B., & Frandsen, T. F. (2018). The impact of patient, intervention, comparison, outcome (PICO) as a search strategy tool on literature search quality: a systematic review. *Journal of the Medical Library Association: JMLA*, 106(4), 420.
- Fabiano, N., Gupta, A., Bhambra, N., Luu, B., Wong, S., Maaz, M., F Fiedorowicz, J. G., Smith, A. L & Solmi, M. (2024). How to optimize the systematic review process using AI tools. *JCPP advances*, 4(2), e12234.
- Grandini, M., Bagli, E. and Visani, G., 2020. Metrics for multi-class classification: an overview. *arXiv*. <https://doi.org/10.48550/arXiv.2008.05756>.
- Higgins J.P., Green S., Cochrane handbook for systematic reviews of interventions 4.2. 6 [updated september 2006], Vol. 4, The Cochrane library (2006)
- Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2), 1.
- Hu, G., Muhamad, H., Mustafa, H., & Alhasnawi, M. Y. (2025). Accounting education from 1960 to 2023: a bibliometric review. *International Journal of Evaluation and Research in Education (IJERE)*, 14(2), 849-861.
- Kitchenham, B., Brereton, P., Li, Z., Budgen, D., & Burn, A. (2011, April). Repeatability of systematic literature reviews. In *15th Annual Conference on Evaluation & Assessment in Software Engineering (EASE 2011)* (pp. 46-55). Stevenage UK: IET.

- Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering—a systematic literature review. *Information and software technology*, 51(1), 7-15.
- Lau, O., & Golder, S. (2025). Comparison of Elicit AI and Traditional Literature Searching in Systematic Reviews using Four Case Studies. *medRxiv*, 2025-06.
- Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M.I., Noel-Storr, A., Rader, T., Shokraneh, F., Thomas, J. & Wieland, L.S. (2019). Searching for and selecting studies. *Cochrane handbook for systematic reviews of interventions*, pp.67-107.
- Owen, B. M., Tweedley, J. R., Moheimani, N. R., Hallett, C. S., Cosgrove, J. J., & Silberstein, L. P. (2026). What is “accuracy”? Rethinking machine learning classifier performance metrics for highly imbalanced, high variance, zero-inflated species count data. *Limnology and Oceanography: Methods*, 24(2), e70009.
- Paynter, R. A. (2009). Evidence-based research in the applied social sciences. *Reference Services Review*, 37(4), 435-450.
- Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.
- Rebele, J. E., Apostolou, B. A., Buckless, F. A., Hassell, J. M., Paquette, L. R., & Stout, D. E. (1998a). Accounting education literature review (1991–1997), part I: Curriculum and instructional approaches. *Journal of Accounting Education*, 16(1), 1-51.
- Rebele, J. E., Apostolou, B. A., Buckless, F. A., Hassell, J. M., Paquette, L. R., & Stout, D. E. (1998b). Accounting education literature review (1991–1997), part II: students, educational technology, assessment and faculty issues. *Journal of Accounting Education*, 16(2), 179-245.
- Rebele, J. E., Stout, D. E., & Hassell, J. M. (1991). A review of empirical research in accounting education: 1985-1991. *Journal of Accounting Education*, 9(2), 167–231.
- Rebele, J. E., & Tiller, M. G. (1986). Empirical research in accounting education: A review and evaluation. In A. C. Bishop, E. K. St Pierre, and R. L. Benke (Eds.), *Research in accounting education* (pp. 1-54). Harrisonburg, VA: Center for Research in Accounting Education, James Madison University
- Stone, P. W. (2002). Popping the (PICO) question in research and evidence-based practice. *Applied nursing research*, 15(3), 197-198.
- van Dijk, S. H., Brusse-Keizer, M. G., Bucsán, C. C., van der Palen, J., Doggen, C. J., & Lenferink, A. (2023). Artificial intelligence in systematic reviews: promising when appropriately used. *BMJ open*, 13(7), e072254.

Van Rijsbergen, C. (1979, September). Information retrieval: theory and practice. In *Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems* (Vol. 79, pp. 1-14). Oxford, UK: Butterworth-Heinemann.

Watson, S. F., Apostolou, B., Hassell, J. M., & Webber, S. A. (2003). Accounting education literature review (2000–2002). *Journal of Accounting Education*, 21(4), 267-325.

Watson, S. F., Apostolou, B., Hassell, J. M., & Webber, S. A. (2007). Accounting education literature review (2003–2005). *Journal of Accounting Education*, 25(1-2), 1-58.