JUTLP
Journal of University Teaching & Learning Practice

Systematic Literature Reviews: Why I Rejected Your Review

Dr Joseph Crawford,
[a] University of Tasmania, Australia

Abstract

Systematic literature reviews (SLRs) have become an increasingly common methodological approach in higher education research, particularly within the Journal of University Teaching and Learning Practice (JUTLP). However, despite their popularity, a significant proportion of submitted SLR manuscripts are desk-rejected prior to peer review. This Commentary critically examines recurring methodological shortcomings that often undermine the credibility and rigour of these submissions. Drawing on best practices in evidence synthesis, this study highlights five core areas where authors frequently fall short: (1) the formulation of research questions that are appropriately scoped, answerable, and aligned with review goals; (2) the development of transparent, valid, and replicable search strategies using Boolean logic, truncation, and multiple databases; (3) the implementation of systematic screening and selection processes, including use of PRISMA flow diagrams and clear inclusion/exclusion criteria; (4) the use of trustworthy and replicable methods of data extraction and synthesis, including quality appraisal of included studies; and (5) the articulation of meaningful implications that extend beyond descriptive summaries to offer theoretical, empirical, and practical contributions for an international readership. Through these insights, this Commentary seeks to provide constructive guidance for researchers and reviewers, with the aim of enhancing methodological integrity and increasing the acceptance rate of SLR submissions. By strengthening methodological transparency, reliability, and relevance, SLRs can serve as powerful tools to synthesise evidence, guide pedagogical innovation, and inform higher education policy.

Editors

Section: Educational Psychology
Senior Editors: Dr Sally-Ann Ashton-Hay, A/Prof Rebekkah Middleton, Dr Cassandra Colvin, and Dr Averil Grieve.

Publication

Submission: 26 May 2025
Revised: 29 May 2025
Accepted: 30 May 2025
Online First: 30 May 2025

Practitioner Notes

1. Systematic reviews are increasingly popular but commonly rejected.
2. Systematic literature reviews should start with a clear and compelling research question.
3. Authors must adopt a valid and replicable search strategy.
4. Systematic literature reviews need to carefully consider initial and full-text screening decisions.

Keywords

Systematic literature review, evidence synthesis, PRISMA, higher education research, methodological rigour, research methodology, desk rejection, search strategy, review quality

Citation:
Crawford, J. (2025). Systematic Literature Reviews: Why I Rejected Your Review. *Journal of University Teaching and Learning Practice, 22*(2). https://doi.org/10.53761/10vb5076

Introduction

The volume of systematic literature review submissions to the Journal of University Teaching and Learning Practice (JUTLP) has risen considerably in recent years. *Prima facie,* this may seem strange for a practice-based journal, especially considering most studies published in JUTLP are empirical assessments of some learning and teaching innovation or practice in local and global contexts. The business of a journal is never to dictate methods, but rather to assess the relevance of a manuscript to the aims and scope the journal sets for its contribution to knowledge. JUTLP Editors assess manuscripts for disseminating improvements to learning and teaching practice and it is difficult for purely theoretical studies to establish a direct connection to practice. However, as Editor-in-Chief, it has been my view that well-designed systematic literature reviews, and meta-analyses oriented around practices and practical outcomes, are well suited to achieving change in international learning and teaching.

Systematic literature reviews offer a mechanism to organise what is known about the practical matters of curriculum, assessment, student experience, education leadership and education technologies in ways that are replicable, valid, and clear. Yet, and increasingly, published reviews often present a description of knowledge contained in the sample of studies as based on a robust process *because* they are reviews, despite being poorly framed or methodologically flawed. As a scholar, I appreciate clarity, rigour and transparency in a systematic literature review. In this paper, I outline the elements I look for that define a rigorous systematic literature review. I set out the key concepts and decision processes that lead to well-designed systematic literature reviews. I advocate for a focus on designing systematic reviews for impact rather than aiming for a lengthy section that summarises what is currently known and makes recommendations for future research (although this is important too).

For each systematic literature review JUTLP accepts for publication, ten to twenty are rejected due to fundamental design issues, including poor search strategies, transparent screening protocol, poor analysis and low impact conclusions and recommendations. This Commentary is a synthesis of the advice I provide authors of systematic literature reviews whose papers I rejected within minutes of reading their methods section. The design advice is distilled from the hundreds of rejection letters I have written and, because I appreciate how dense discussing methodology can be, I use song lyrics at each of the signposted points.

Robust Research Questions

Oasis, while not the first band to ask, "*What's your story, morning glory?"* follows with "*I need a little time to wake up, wake up*". I find this apt for designing good systematic literature reviews. The reader of the review needs a clear understanding of why this work is relevant in the first place and what problem it seeks to solve. A clear problem statement should be framed inside of practice and research. Too often, however, problem statements in systematic reviews are often left blank or include tautological arguments like 'not enough is known about this field' or 'the knowledge is disparate and needs organising'. While these have some general merit, it is the role of the researcher to articulate that a field of inquiry is worthy of the effort required to do a literature review. I can imagine that little is known about the link between the personality traits of sheep and their propensity to complete tertiary qualifications, but I lack the interest in exploring that topic systematically. In Crawford et al. (2024), we test a critical appraisal tool for theoretical and literature-based studies – the Quality Assessment Tool for Theory-Based and Literature Review

Studies (QATTL) – and propose this statement to measure the robustness of a systematic literature review's research question: *the research question(s) are justified and clear.* The IDEA-ARC model for developing and communicating research questions is another useful framework for assuring a good research question (see Purvis et al., 2024). In the 'IDEA' stage, Purvis and colleagues encourage researchers to: Identify a clear problem; Draft background information; Explore diverse approaches to addressing the question; and Develop agreement on a question. The 'ARC' stage includes: Applying the question; Reflecting on the impact of the research on addressing the question; and Communicating the outcomes of the original questions. In this next section, I focus primarily on the IDEA stage from a practice and a research perspective.

Practical problems

Practical problems (e.g., a group of students are struggling to belong) are harder to define than theoretical problems (e.g., we lack insight into student belonging, generally), particularly when seeking practical knowledge from theoretical studies such as a systematic literature review. However, a systematic literature review can offer an effective pathway by critically consolidating the literature guided by an underlying aim to inform and improve practice.

Take the daily experience of a Deputy Vice Chancellor (Education) in leading the policy, strategy, and practice of learning and teaching in their university. They have limited time and often find that most published studies on a given topic are single institutional studies or practice reflections with limited assurance of successful application to their context. So, instead of adopting new practices, they opt for a more conservative option. That is until a well-crafted systematic literature review with clearly defined focus comes across their desk. In an hour or two they have caught up with the latest literature in an organised fashion and can see a myriad of opportunities where the review findings can be applied into their context. While systematic literature reviews can advance disciplinary fields and theory, they are most useful for JUTLP if they are clearly situated to address a problem in practice.

Good systematic reviews require positioning and clarity to clearly frame the search strategy to come. When designing a systematic literature review, it can be easy to conceive a generally targeted audience such that, *this paper will be useful to educators, policymakers, and higher education leaders.* For example, a systematic literature review on GenAI in learning and teaching – currently a popular submission in JUTLP – can be framed as an educational technology review focused on the technical aspects of GenAI, an educational psychology review emphasising the benefits and consequences of GenAI on the student or educator, or a curriculum review framed to inform GenAI policy. All of these are noble pursuits if clearly defined, but 'GenAI in learning and teaching' is too broad to be meaningful to future practice. A large-scale review may only serve as a catalogue of studies, rather than an attempt to influence the innovation of practice.

In the same line, to offer a generalist articulation of a wicked problem or broad issue without clearly explaining the conceptual justification for it being the focus of a review is unlikely to be materially useful. I frequently reject systematic reviews that describe a general issue such as an 'increase in student cheating' and argue, therefore, a systematic literature review is the solution. Instead, the problem formulation could be distilled further to constitute a more specific and applied examination of that broad concern. For example: a review of student behavioural intentions to cheat; motivations to cheat; assessments more prone to cheating; character and cheating; detection techniques; contract cheating access; or institutional response strategies. The general formulation does not offer the level of specificity required to design an effective search strategy, which can lead to a well organised corpus of knowledge aligned to the review's purpose that can inform changes to practice. Identifying subsets of the core issue – e.g., student cheating – offers a more tangible set of solutions for educators or policy makers to specifically tackle a subset of student cheating in their institutions.

Research and theory problems and answers

Strong research – particularly systematic literature reviews – offer a pathway forward for practical problems and extend to clearly evidencing a research agenda for the future. A key benefit of a review is to identify what is known, where the knowledge is contested, and where there are gaps in knowledge. While the use of the large-scale claims of single sample studies using a local dataset without replication or re-testing is typically considered academic overreach, the systematic literature review identifies a sample of associated studies and generates claims using the whole corpus.

Claims arising from a review can be as simple as stating consistent issues of the field. For example, Day and colleagues (2024) comment in a 25-year review on leadership development that "a significant obstacle to advancing scholarly interest in leader and leadership development over the years can be traced to methodological and analytical issues" (p. 77) and go onto highlight key issues and responses related to methodology. They offer a research proposition for improving the future of their concept – leadership development – alongside offering a review of the underlying theories driving the field. Similarly, in a recent systematic literature review of teacher pedagogical competencies, Moreira et al. (2023) argue that the competency of

> Collaboration is still present [in the research], although not highly valued for defining quality for the teaching profession in this context, contrary to what is advocated for non-tertiary teaching. Therefore, teaching in higher education seems to be a more solitary professional activity than teaching at other education levels (p. 111).

A systematic literature review should highlight what is established as known about the phenomenon of interest and, as much as possible, offer a consolidated description and evaluation of the concept. For example, in a recent review, Hattie and O'Leary (2025) examine 17 meta-analyses that evaluate the relationship between achievement and learning styles. By drawing systematically on this abbreviated sample, their review highlights seven research confounders and raises an important question, "... why does the belief in the discredited concept of learning styles persist?" (p. 31). Given their claim based on the analysis, they then offer some alternatives, that

> The persistence of learning styles as a concept in educational discourse and research is paradoxical, given the overwhelming evidence discrediting the matching hypothesis, the notion that aligning teaching methods with students' preferred learning styles enhances achievement (p. 31).

Authors of systematic literature reviews can pay lip service to the goal of advancing knowledge through future research agendas by generating star-sign styled statements that will generally be true like 'more research is needed'. For example, a recent review published on ChatGPT and student engagement offers an important summary of how this specific generative AI tool is used to enhance and hinder student engagement (Lo et al., 2024). The review also promises a future research agenda, but falls short of meeting quality criterion (e.g., QATTL) in practice. Starting with a clear research and theoretical problem allows big questions to be asked of future research or perhaps to be answered in the context of the review itself.

Search strategy

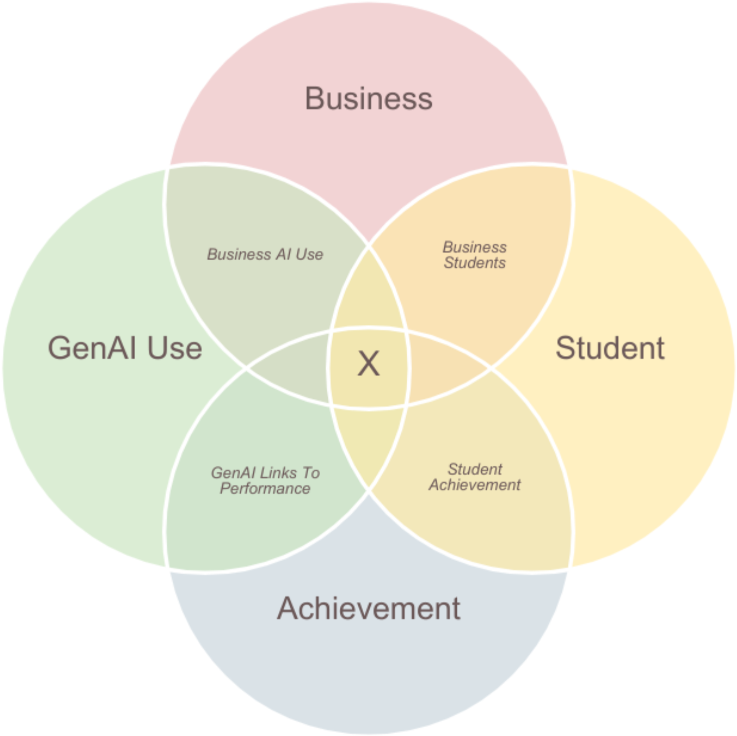
The endpoint of a process of comprehensive literature reading and reviewing with a focus on a practice-relevant topic is hopefully a research question, which is evidence that "*I still haven't found what I'm looking for"* (U2). That is, the attempt to disprove the need for a systematic literature review on the specific topic failed. The next step is to design a search strategy. I argue that search strategy is the most important step to do correctly; failure at the search strategy stage almost invariably requires restarting the full review process. I empathise with authors when I send a rejection letter pointing out there is a search term they have not included in their initial search strategy or a Boolean configuration that misses a simple variation to their terms. For this section, I use a worked example (designed to be problematic) based on the following question: *What is known about business students' use of generative AI tools and their academic achievement?*

Develop robust search phrases

The robustness of a systematic literature review depends in part on its capacity to assure and validate that all the relevant publications that ought to be considered are included in the initial results. I have observed many approaches to effectively determining the fields of inquiry and attending to their intersections. In Figure 1, I identify four key frames associated with the example question. The goal of a search strategy is to find those studies that meet the middle point X, or those studies that have something in them that relates to all elements of the broader search. However, using these search term words alone in a search (e.g., business AND student AND GenAI use AND achievement) is not likely to yield a robust search outcome – the reasoning for which is simple: not all papers use the exact terms that we wish they did. For example, if the study was on accounting students and never referred to business in the abstract, it would be ineligible despite its relevance so the search needs to be extrapolated.

Figure 1

Search strategy Venn diagram



Where possible, it can be useful to rely on historical searches that have been successful. In Table 1, I identify three similar searches and their differences and make a recommendation for the search I would have conducted. To further evidence this, I run a title or abstract search on Web of Science English articles with a data range from 2000-2024.

Table 1

Student Search Phrases

Citation	Search phrase	Results
Arulkadacham et al. (2022)	“tertiary” OR “higher education” OR “university” OR “college”	943,207
Blair and van der Sluis (2022)	“university” OR “college” OR “higher education” OR “HE” OR “post-compulsory education”	1,003,695
Fadlemula and Qadhi (2024)	“higher education”, “university”, “undergrad”, “college”, “tertiary” OR “post-secondary education”	943,955
Recommended search	“higher education” OR university OR undergrad* OR postgrad* OR college OR tertiary OR post-secondary	1,039,726

There is variance in the search term selection, and there are other important distinctions in relation to phrase search, Boolean operators, truncation, and proximity.

Table 2

Search term variations

Term	Definition	Example
Phrase search	The ability to search for individual words or specific sequences of words.	"higher education" versus higher education yield different results.
Boolean operators	Use of logic parameters to help refine to what output is desired.	AND = terms on the left and right are both required. OR = either term on left or right are required. NOT = the term following NOT should not be present in the papers.
Truncation	Wildcard options available to select partial words.	* = specifies a root word that has variation. *student** will allow for student or students in the search. ? = allows for the replacement of a single letter. Organi?ation may allow for US and UK spelling. $ = allows replacement of zero or one character. Colo$r allows for US and UK spelling.
Proximity	Allows for two words to be selected within a certain closeness of each other.	NEAR/x = searches for two words within x words of each other, in either order. Student NEAR/3 engagement = searches for any combination of student engagement within three words, such as "engagement of students".

The goal of a good search phrase is to ensure all the necessary combinations are included by clearly defining the levels of a search. There are robust examples of multi-frame searches (e.g., Allen et al., 2024). To develop a search phrase for '*What is known about business students use of generative AI tools and their academic achievement?'* the goal is to be comprehensive. The process I recommend ensuring the robustness of the search terms is iterative searches. The first iteration involves running a sample test and reviewing possible results, such as running an English article Web of Science search and reading abstracts until saturation is met, including looking for keywords in those abstracts that may add depth to the subsequent searches. For example, after noting that a paper on management students classifies some individuals as leadership students, adding 'leadership' to the first search phrase will ensure that related studies are included. For the purposes of this example, I am using an adapted multi-frame search derived from Broadbent and Poon (2015) for academic achievement and Law (2024) for GenAI:

> (Management OR accounting OR finance OR business OR economic* OR marketing) AND ("higher education" OR university OR undergrad* OR postgrad* OR college OR tertiary OR post-secondary) AND ("academic outcome" OR "academic attainment" OR "academic accomplishment" OR achievement OR score OR mark* OR rank* OR GPA OR grade* OR success OR performance) AND ("generative artificial intelligence" OR "generative AI" OR "GenAI" OR "ChatGPT" OR "Chat GPT").

The second step is to run the full search iteratively, deleting one of the search frames for each iteration. For example, including the first three frames from above, and iterating to only include "generative artificial intelligence", with a second search using "generative AI" and so on. This allows assurance that each keyword has relevance to the search and yields at least one result. With that, an effectively defined search phrase is articulated and validated. There may likewise be value in considering manual re-searching to ensure no items are missed or, in the event of a micro-sample, reviewing at a journal-level may be helpful (e.g., Ashton-Hay, 2025).

Where should I search? Databases, types, and search selection

With the search phrase in hand, a decision is needed as to which databases are appropriate, which type of documents will be included, and which filters are to be used in refining the search. Some authors choose to include method-related keywords to limit their search, however a significant issue with incorporating method-based keywords into a search is that including method terms in the title or inside the abstract is not consistent, and searching for a method across a full-text will still yield substantial inconsistencies. For example, a paper referring to an experiment in their literature would be included in a search requiring experimental methodological language.

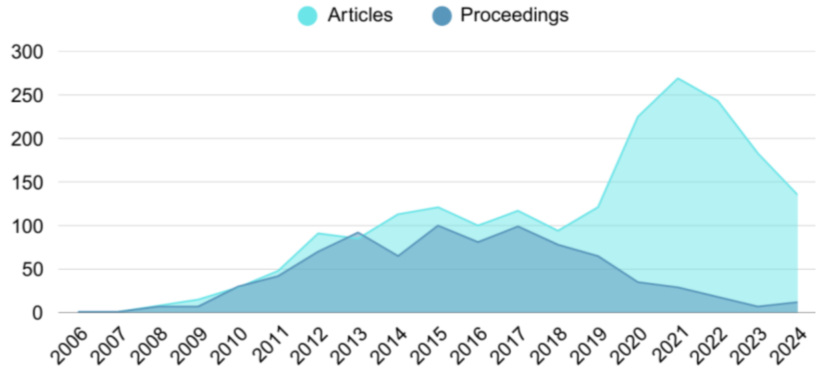
On databases, it is useful to consider the database section reported in a systematic literature review published in a top journal in the field. A reuse of a well-selected suite of databases is an efficient practice, particularly in citing the original source. Web of Science and Scopus appear to be now fundamental in all search strategies when journal articles are in mind, given these offer the most comprehensive list of quality manuscripts available. In addition to Web of Science and Scopus, discipline specific options are recommended such as PsycInfo and PubMed for health and psychology, ERIC for education, ProQuest for dissertations, and Google Scholar for secondary searching. In higher education and most social sciences, a combination of Web of Science, Scopus, PsycInfo, and PubMed with a secondary manual search in Google Scholar appears to be generally robust (Yang et al., 2025). In the case of manual searches, some authors choose to conduct a manual reading of a specific journal particularly relevant to the topic. However, manual searches are becoming less common as databases continue to build robustness and breadth.

Selection of document types such as peer-reviewed research articles, conference proceeding papers, book chapters, datasets, and grey literature, is an important decision point for authors. Peer reviewed article-only systematic literature reviews are common, given the quality of a peer reviewed paper can be generally assured; noting that retracted papers, withdrawn articles, publications with expressions of concern, and those where peer review is not assured should be removed from the sample. The aspiration in selecting document types is to consider the rigour of each study that has been identified by the search terms as likely to address the research question. Although the least reliable, grey literature may be useful if the aim is to review consumer or community sentiment. Conference papers may be considered for emergent topics where much of the literature may not have yet been published in top tier journals (e.g., scoping reviews on ChatGPT in early 2023). Where a topic already has considerable depth, theses and dissertations may also add value.

Year range requires careful justification and design. In the selection of years, a range that is too large can make for an impossible task for the research team. Alternatively, a year range that is set a year too short could make a study redundant. To decide on the appropriate timeframe, a practical logic is useful. For example, Figure 2 shows the results by year of a search for the higher education search frame in Table 1 AND ‘Facebook’ and including articles and proceedings. In the first 10 years of studies since the release of Facebook, the high proportion of conference papers makes sense, especially given practice and innovation dissemination in conferences is high. After 2019, with increasing maturity of the field, conference papers decline, and the volume of peer reviewed articles rises and then falls. Studies relating to sustainable development goals as released by the United Nations practically begin their search from the day of the Millennium Development Goals on 1 January 2016, and those involving questions about the Bologna Declaration may be linked to the 1999 date of signing.

Figure 2

Articles and proceedings papers on Facebook in higher education



Snowball sampling or chain-referral-sampling is a supplementary practice that draws on non-literature review sampling techniques to assure papers related to a final sample are included. In this method, authors review by title alone the reference lists of the final papers to be included to check for any possible papers that may be worthy of considering for analysis. This method can offer a mechanism to further quality assure the original search – if many articles are identified it signals that at least one keyword is missing from the search terms. Snowballing is most useful in contexts where access to the right studies is difficult to locate, such as when keywords are naturally ambiguous (e.g., reviewing a broad-concept curriculum change).

Is it enough papers?

At the conclusion of the search strategy, the research team should have a large corpus of manuscripts far greater than the volume of studies expected for final inclusion. It is common for scholars at this stage to ask, "Is it enough?" Indeed, I ponder this with frequency. How many papers are required for a robust systematic literature review depends highly on the maturity of the research topic and the constraints placed on the search strategy. Systematic literature reviews can be conducted by individuals who are new to the field (i.e., developing expertise through review) or highly experienced (i.e., reflecting on the field with a systematic approach). The latter individuals likely have a good sense of the extent of research currently on a topic, with refined research questions likely to lead to narrow results and low volume. The former need to bridge the gap in their understanding of the field to support a robust assessment of whether the search strategy has yielded sufficient possible papers to lend itself to an effective review. Having now published, edited, and reviewed systematic literature reviews with small (< 10) and enormous (> 1,000) samples, in my judgement, a corpus of final papers less than 20 is too small to yield the kinds of systematic analyses that make a study interesting and insightful to the field.

For meta-analyses, McLennan and Perera (2018) argue that analyses using a subset of nine studies achieves the same output as the full sample (80% of the time), and Cuijpers et al. (2021) suggest that only six studies are required for 80% power. An important distinction in these analyses relate to whether subgroup analyses are planned, where 22 is argued as a robust minimum (Cuijpers et al., 2021). That is, whether the analyses are planned to be carved out among specific types of participants (e.g., gender, age, experience) or outcomes (e.g., separating loneliness studies by UCLA and other measurements). Planned systematic literature reviews that comprise less than 20 research papers may be better suited to considering scoping review methodologies to take the study beyond the screening stage and offer useful insights into a nascent and emergent field. A practical screening heuristic may be to consider at minimum of at least three papers per year of searching (e.g., a 10-year review of no less than 30 papers) to offer sufficient density and temporal relevance to make a meaningful contribution to the literature.

Initial and Full-Text Screening Decisions

The search is finished and now there is a large volume of unscreened titles and abstracts uploaded into one of the many software options available to manage the screening process of a systematic literature review. The next step is to make justified decisions about managing the screening, handling conflicts, and considering artificial intelligence applications that now support this process.

Initial title and abstract screening and secondary full-text screening

Most systematic literature reviews that adopt a Preferred Reporting Items for Systematic Review and Meta Analysis Protocols (PRISMA-P: Moher et al., 2015) or other variations, incorporate two phases of screening. The initial title and abstract screening is designed to focused on efficiently culling irrelevant manuscripts based on their titles and abstracts. The process is particularly useful when some of the keywords used can have known but confounding meanings. For example, in a systematic literature review of lectures in universities, a large volume of manuscripts were found to be transcripts of distinguished and invited lectures with ‘lecture’ in their titles (Crawford & Parsell, 2025). Similarly, in an in-progress meta-analytic review on social connection at work, there are many confounding manuscripts that discuss two elements in workplaces being connected through a quantitative analysis, but the papers are not actually related to human or social connection (Chanko et al., 2025).

The screening process allows for researchers to make judgements on which manuscripts are relevant and acceptable to complete a full review, and which are clearly irrelevant. At this stage, a simple version of the inclusion criteria is applied. For the example business student review, inclusion may be as straightforward as ensuring it meets the criteria of English, peer-reviewed journal, relates to some kind of business student, makes a reference to their academic performance, and includes a reference to generative artificial intelligence. Importantly, many manuscripts will meet this check without it being clear whether the manuscript will meet the full inclusion criteria. At this stage, it is generally okay to include the manuscript for second screening, recognising that its exact appropriateness will be determined in a full reading. For example, the abstract may reference student academic achievement, without it being clearly linked to the business student's GenAI use. It could simply be that the abstract was not written for the purpose the review intended, so a full reading at the next stage is advisable.

The full-text screening provides a secondary opportunity to consider the manuscripts that look appropriate from a title and abstract reading and ensure that each manuscript meets the full inclusion criteria. In the example review, the manuscript must draw a clear link in its findings between how business students' use of GenAI is influencing their performance. If GenAI was part of a manuscript's literature review and discussion but was not a component of the study's data collection and insight, its relevance to the review is only as a secondary citation. Referenced papers in their literature review may be relevant studies and, if the search strategy was robust, will appear in the final sample.

Transparency and integrity in the process of full-text screening is key. In manuscripts submitted to JUTLP, it is not uncommon to see that authors report they concluded their search at the end of the previous year, and have a full manuscript submitted in February or March despite stating that the authorship team of three read through 200 full manuscripts in the full-text screening. Editors and reviewers are likely to view this with scepticism and require justification. For example, it would be more feasible if authors had screened papers prior to and during 2024 and completed an updating search in early 2025 prior to completing the full analysis and write-up. At the full-text stage, there should be no 'maybes' included; the authors should be confident that every manuscript that progresses to analysis has strong relevance and meets the inclusion criteria.

Generating reliable inclusions: Interrater and intrarater reliability, conflicts, and screening teams

Decisions of who ought to be involved in the screening vary across research projects. The aspiration of systematic literature reviews is to be replicable and ensure that individual error or bias is minimised as much as is practicable. In a review of systematic literature review methods for screening, Waffenschmidt et al. (2019) finds that there is not equivalency between single- and double-screened reviews with the former missing substantially more papers (median proportion of missed papers ranged between 0.6% to 16.6%: median 5%). On face value, this may indicate that single reviews should not be undertaken, but this is not the only possible outcome. Reviewer experience and a clear research question was found to be a key distinguishing factor for successful (<5% error rate) versus unsuccessful single screenings. The decision factor in many studies is balancing workload/funding expectations and robustness of outcomes (Shemilt et al., 2016).

In a large study (15,000 abstracts) using a second reviewer for study selection, Stoll and colleagues (2019) identified an additional 4.4-5.3% of eligible papers were included (6.6-9.1% in initial screening, and 6.6-11.9% in full screening). Where funding and resources do not permit a dual screened initial and full-text screening, an approach of sample retesting may be useful, although not a direct replacement for two reviewers across the full sample. In this approach, authors may have a single reviewer but seek a second reviewer on a small proportion of manuscripts (e.g., 5-20%) and conduct interrater reliability testing on this proportion prior to completing the final sample as individual(s). The single reviewer could also introduce a delay between a first screening and complete an intra-coder retest equivalent to approaches taken in qualitative coding (Roberts et al., 2019). These options are presented in Table 3 in order of reliability.

Table 3

Types of screening

Screening Type	Definition	Reliability Test
Single screening	Reviewer 1 screens all manuscripts without any retesting.	None
Single screening with time-lagged intrarater retest	Reviewer 1 evaluates some manuscripts and inserts a time-lagged blind re-evaluation of the initial sample and compares consistency.	Cohen’s kappa Percent agreement Intraclass correlation
Single screening with subset interrater retest	Reviewer 1 evaluates some manuscripts and Reviewer 2 completes a blind evaluation and compares consistency.	Cohen’s kappa Percent agreement Gwet AC1/AC2
Double screening	Reviewers 1 and 2 evaluate all manuscripts independently and compare consistency.	Cohen’s kappa Fleiss’ kappa* Percent agreement Gwet AC1/AC2

*When more than 2 reviewers are used.

Cohen’s kappa and percent agreement are reasonably common methods for detecting reliability, with the former ideally seeing scores in the moderate (.60-.79), strong (.80-.90), and almost perfect (>.90) categories (Landis & Koch, 1977; McHugh, 2012). Percentage agreement is simple to calculate, and there are a good deal of Cohen’s κ calculators online that accept an input of raw scores and calculations (e.g., Table 4). The Fleiss kappa is more useful when a team of reviewers are interchangeably taking on roles of Reviewer 1 or Reviewer 2, i.e. all manuscripts have two reviewers, but may comprise different configurations of the team pending how software organises them. Gisev et al. (2013) provide a useful worked example of calculating this score. Of relevance to the decision-making process is considering temporal points for assessing reliability. From observation, this often occurs at the end of the process as a confirmatory measure. That is, simply reporting the Cohen’s Kappa was satisfactory at the end is suitable if the data supports it, and it can also be used as a formative measure midway through the analysis to pause, assess, and make corrections to improve collective agreement. In Table 4, I offer a simple example with mock data to highlight calculation methods with a 1 for recommended inclusion, and 0 referencing a recommended exclusion.

Table 4

Sample manuscripts for reliability calculations

Manuscript ID	John	Ben	Difference
1	1	0	1
2	1	0	1
3	1	1	0
4	0	0	0
5	1	1	0
Zeros			3
Manuscript *n*			5
% Agreement	(zeros / *n*) =		60%

	Ben (Y)	Ben (N)	Total
John (Y)	2	2	4
John (N)	0	1	1
Total	2	3	5

Should artificial intelligence do the screening?

In recent years, the emergence of robust machine learning technologies has resulted in a swathe of tools designed to automate and semi-automate the title, abstract, and full-text screening processes. A series of comparisons of some automated screening has been conducted. For example, in one study, Research Screener is used to rank manuscripts and provide a shortlist of likely papers, identifying that only the top 35% of manuscripts (less in some reviews) needed to be reviewed prior to all manually identified papers being included (Chai et al. 2021). In another study, an R/Python tool was used with an outcome of a 61% decrease in workload with a five percent false negative rate (Kebede et al., 2023). There are opportunities to engage in the use of artificial intelligence in the screening process, and usually this is limited to a semi-automatic process where the tools are used with reasonable scholarly oversight. That is, the authors must be confident that if a similar scholar to themselves – or even themselves – replicated the screening process manually that the final included sample would be the same (or ~95% the same). In the absence of this level of assurance of equivalence of screening processes, scholars ought to avoid the integration of machine learning in their screening process.

Analysis Decisions

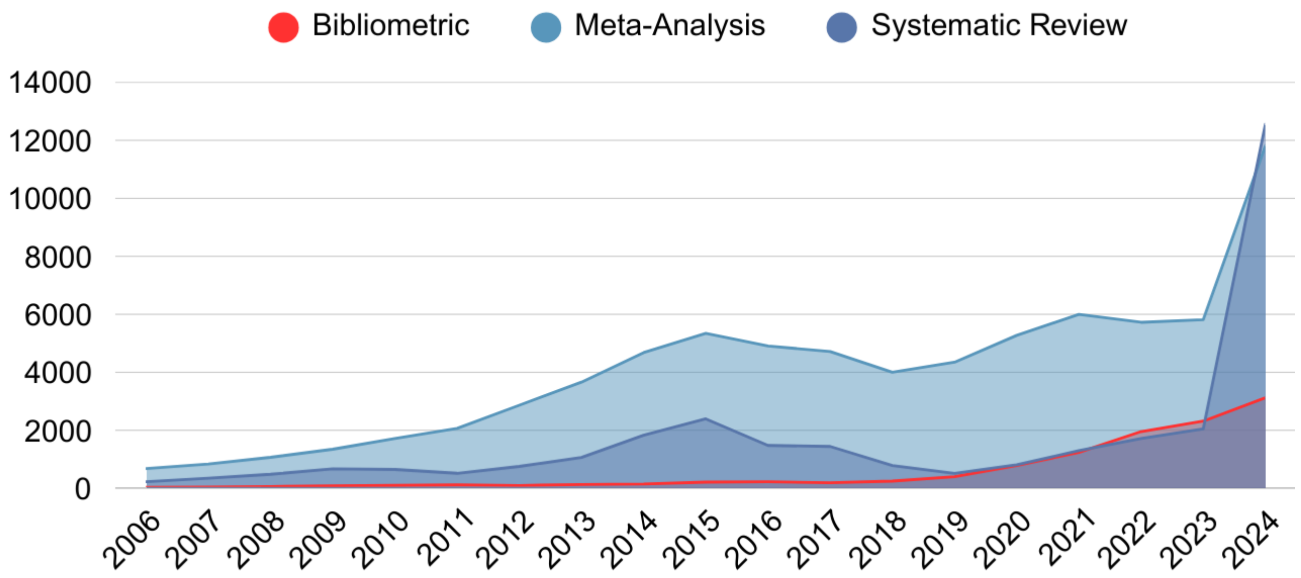
The decision following data collection and filtering is to determine methods for analysis. Ideally, the decision for analysis will be made upfront to enable effective search strategies to be informed by the kinds of analysis proposed. In this section, I highlight two common methods – bibliometrics and thematic analysis – and specifically exclude meta-analysis discussion. Meta-analyses adopt a systematic literature review front-end, but the analysis is too sophisticated to discuss with fairness within the scope of this Commentary. Much of the discussion around thematic analysis may have application to broader qualitative analyses of the final sample (e.g., a narrative or content analysis). I also exclude explicit discussion on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA: Moher et al., 2009; Page et al., 2021). PRISMA is an expectation most authors of submissions to JUTLP respond to, at least in the context of presenting a figure highlighting sample from first search to final inclusion. However, I note that often authors lack transparency on the actual items of reporting.

Bibliometric Analysis

Bibliometrics is a burgeoning field, typically publishing more in recent years compared to systematic literature reviews. Figure 3 highlights a brief search at the title level for “bibliometric”, “meta-analysis OR meta-analytic”, and “systematic review OR systematic literature review”. Aside from significant increases in 2024 (possibly an artificial intelligence inflation), bibliometric reviews in recent years have been closing the gap with meta-analyses and have outpaced the rate of growth of systematic literature reviews.

Figure 3.

Comparing titles for bibliometric, meta-analytic, and systematic reviews



Although their meaning and applications are often misused or misleading, the purpose of a bibliometric analysis is to examine the literature for trends:

> Bibliometric analysis is a popular and rigorous method for exploring and analyzing large volumes of scientific data. It enables us to unpack the evolutionary nuances of a specific field, while shedding light on the emerging areas in that field (Donthu et al. 2021, p. 285).

Bibliometrics are often conducted to understand a particular field – such as authors, citation profiles, demographics of authors or samples, and locations of publication. The goal is often to make sense of possible gaps in the types of analysis conducted. For example, Crawford (2025) highlights gaps in Australian higher education research in the top twenty publications and points to a series of journals that seem to heavily underrepresent Australian-based research. Similarly, Madden et al. (2020) examine gender distribution across medical education editorships and authorships. These are useful for progressing change in areas where there might be deficiencies in the types of research being undertaken, but do not always produce meaningful findings. For example, summarising the number of leadership theories used over a decade may be a useful introductory remark, but would be much stronger as an integrated framing to a literature review. JUTLP rarely publishes bibliometric analyses as, without careful framing and design, they do not tend to change end-user practice. Their goal is to change practice in the process of publishing through enhanced visibility.

Thematic Analysis

A deep analysis of samples in systematic literature reviews often lends itself to the application of text-based qualitative analyses from thematic analysis to phenomenological analyses. For many, this might mean a reference to the original Braun and Clarke (2006) six-step model of thematic analysis from familiarising oneself with the data; generating initial codes, searching for themes, reviewing themes, defining and naming themes, and producing the report. Given there are multiple subsequent publications by Braun and Clarke, it is essential to differentiate the approach of thematic analysis used, including variations on the six-step process published in 2006. In their more recent 2023 study, Braun and Clarke (2023) highlight ten woes of contemporary thematic analysis deployment, which are worthy of consideration. I will summarise those that relate more closely to its deployment in systematic literature reviews. Many systematic literature reviews seem to mischaracterise Braun and Clarke's approach to thematic analysis, declaring their close deployment of the six-steps in their work but producing a descriptive account of themes that appear incoherent and incomplete. That is, their thematic write-up is a basic description of papers included in the sample without critical synthesis and evaluative judgment. Extending on this, many fail to locate themselves qualitatively and instead preference a quantitative style of code reference counting and reporting. While presenting distribution of references may be useful to readers to understand the prevalence of themes, it is insufficient to produce a clear view of themes evident in the sample of published literature. An essential addition to this Commentary is highlighting the lack of transparency common among systematic literature reviews submitted as to what was *actually* done for the thematic analysis. Was it deductive or inductive? What process was taken to be reflexive and iterative? Is there a clear attempt to withdraw (or disclose for those adopting insider perspectives) themselves where possible from the materials to be robust? Each thematic analysis process is different and key to believing the trustworthiness or reliability of the work is transparency in describing and justifying the approach taken.

The future you will change – practice, research, *and* theory

The role of systematic reviews is likely to come into question as their production has accelerated in the last twelve months and most probably will continue to increase as machine learning reduces the workload associated with the overly burdensome process of screening. My concern is that many of the reviews coming across my desk that have deployed the systematic process robustly fail to 'move the dial'. They have produced a review that summarises the literature to date but expects the reader to identify the gaps in knowledge, despite the fact that the authors, whose research involved them deeply canvassing the field are best placed to start that conversation. Increasingly, the scholars whose reviews go on to be significant take the opportunity to make bold and robust claims about the field. That is, a good review should make advances in practice, research, and theory (Shepherd & Suddaby, 2017). Table 5 provides brief summaries of these.

Table 5

Summary of advances systematic reviews should make

Concept	Definition	Example
Theory	To advance knowledge of the underlying phenomena.	Offering a theoretical revision or exposing a now-redundant element of theory, like a call to abandon learning styles.
Research	To advance knowledge of the necessary research methods and approaches needed for theory and practice.	Offering clarity on limitations and strengths of approaches to research methods, such as a call for less self-report surveys.
Practice	To advance the practices of the phenomena under investigation	Providing clear advice to practitioners who are using the current practice, like educators deploying a particular pedagogy.

Importantly, when the review is completed, it can be useful to circulate a working manuscript within a researcher's network of scholars who regularly use the theory under investigation. It can be useful to solicit advice from a researcher perspective for initial critique in hopes of a more refined set of propositions that aims to advance theory, research, and practice. This informal peer consultation also aims to provide a pre-emptive response to possible editor or reviewer critique.

A common recommendation concluding a systematic literature review manuscript – of which I am also guilty – is to recommend more advanced research methods be used. For example, Lo and colleagues (2024) encourage more experimental designs and objective measures, and recommend that future research evaluate evolving capabilities of the chatbot using mixed-methods research. These are useful suggestions for future researchers but miss the opportunity to leverage a strong systematic literature review output to position future research. Hackman and Wageman (2007) offer a strong alternative to this approach and, drawing on their long experience, pose five theoretically robust questions designed to move the field forward. Granted, their work would benefit from having been more systematic in its formation, but their five questions have guided and driven future editorial decisions on papers and informed the design of future research questions.

Conclusion

This Commentary aimed to highlight the myriad of ways that scholars can engage in a high-quality systematic literature review. It brings to light issues that are frequent reasons for rejecting a submitted manuscripts to JUTLP and draws on my author and reviewer experience of systematic literature reviews in the fields of business management and higher education. This Commentary offers a series of steps scholars can take to enhance the robustness of the systematic literature review process and increase their publication prospects.

Acknowledgements

The author discloses that he has no actual or perceived conflicts of interest, beyond an Editor-in-Chief role, in the journal in which this Commentary is published. The author acknowledges the advisory support of the Senior Editors and Dr Jo-Anne Kelder in earlier versions of the manuscript. The author discloses that he has not received any funding for this manuscript beyond resourcing for academic time at their respective university. The author has used ChatGPT to support finding useful cases, but did not employ ChatGPT to participate in the write-up of this research in alignment with principles from Crawford et al. (2023).

References

Allen, K. A., Slaten, C., Lan, M., Craig, H., May, F., & Counted, V. (2024). Belonging in higher education: A twenty-year systematic review. *Journal of University Teaching and Learning Practice*, *21*(5), 1-55. https://doi.org/10.53761/s2he6n66

Arulkadacham, L., McKenzie, S., Aziz, Z., Chung, J., & Dyer, K. (2021). General and unique predictors of student success in online courses: A systematic review and focus group. *Journal of University Teaching and Learning Practice*, *18*(8), 1-22. https://doi.org/10.53761/1.18.8.6

Ashton-Hay, S. (2025). Student Voice: Reviewing two decades of the literature to guide the next 20 years. *Journal of University Teaching and Learning Practice*, *22*(1). https://doi.org/10.53761/w9tde483

Blair, E., & van der Sluis, H. (2022). Music performance anxiety and higher education teaching: A systematic literature review. *Journal of University Teaching & Learning Practice, 19*(3). https://doi.org/10.53761/1.19.3.05

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2023). Is thematic analysis used well in health psychology? A critical review of published research, with recommendations for quality practice and reporting. *Health Psychology Review*, *17*(4), 695-718. https://doi.org/10.1080/17437199.2022.2161594

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, *27*, 1-13. https://doi.org/10.1016/j.iheduc.2015.04.007

Butler-Henderson, K., & Crawford, J. (2020). A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity. *Computers & Education*, *159*, 104024. https://doi.org/10.1016/j.compedu.2020.104024

Chai, K., Lines, R., Gucciardi, D., & Ng, L. (2021). Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. *Systematic Reviews*, *10*, 1-13. https://doi.org/10.1186/s13643-021-01635-3

Chanko, A., Crawford, J. & Wickham, M. (2025). Sense of belonging and social connection in office design: A meta-analysis. *Unpublished manuscript.*

Crawford, J., Cowling, M., Ashton-Hay, S., Kelder, J. A., Middleton, R., & Wilson, G. (2023). Artificial intelligence and authorship editor policy: ChatGPT, Bard Bing AI, and beyond. *Journal of University Teaching and Learning Practice, 20*(5). https://doi.org/10.53761/1.20.5.01

Crawford, J., & Parsell, M. (2025). Lectures in higher education: A 22-year systematic review. *Journal of Applied Learning and Teaching*, *8*(1). https://doi.org/10.37074/jalt.2025.8.1.25

Crawford, J. (2025). Australian higher education researcher between 2020-2024: Open-access fees, authorship, editorships, and institutional analysis. *Journal of University Teaching and Learning Practice*, *22*(1), 1-15. https://doi.org/10.53761/4a3v2j43

Cuijpers, P., Griffin, J., & Furukawa, T. (2021). The lack of statistical power of subgroup analyses in meta-analyses: a cautionary note. *Epidemiology and Psychiatric Sciences*, *30*, e78. https://doi.org/10.1017/S2045796021000664

Day, D., Fleenor, J., Atwater, L., Sturm, R., & McKee, R. (2014). Advances in leader and leadership development: A review of 25 years of research and theory. *The Leadership Quarterly, 25*(1), 63-82. https://doi.org/10.1016/j.leaqua.2013.11.004

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Fadlelmula, F., & Qadhi, S. (2024). A systematic review of research on artificial intelligence in higher education: Practice, gaps, and future directions in the GCC. *Journal of University Teaching and Learning Practice, 21*(6), 146-173. https://doi.org/10.53761/pswgbw82

Gisev, N., Bell, J., & Chen, T. (2013). Interrater agreement and interrater reliability: key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *9*(3), 330-338. https://doi.org/10.1016/j.sapharm.2012.04.004

Hattie, J., & O'Leary, T. (2025). Learning Styles, Preferences, or Strategies? An Explanation for the Resurgence of Styles Across Many Meta-analyses. *Educational Psychology Review, 37*(2), 1-26. https://doi.org/10.1007/s10648-025-10002-w

Hackman, J., & Wageman, R. (2007). Asking the right questions about leadership: Discussion and conclusions. *American Psychologist, 62*(1), 43-47. https://doi.org/10.1037/0003-066X.62.1.43

Kebede, M., Le Cornet, C., & Fortner, R. (2023). In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature. *Research Synthesis Methods*, *14*(2), 156-172. https://doi.org/10.1002/jrsm.1589

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174. https://doi.org/10.2307/2529310

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, 100174. https://doi.org/10.1016/j.caeo.2024.100174

Lo, C., Hew, K., & Jong, M. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education*, 105100. https://doi.org/10.1016/j.compedu.2024.105100

Madden, C., O'Malley, R., O'Connor, P., O'Dowd, E., Byrne, D., & Lydon, S. (2020). Gender in authorship and editorship in medical education journals: a bibliometric review. Medical Education, 55(6), 678-688. https://doi.org/10.1111/medu.14427

McHugh, M. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

McLellan, J., & Perera, R. (2018). 21 Restricted meta-analyses versus full meta-analyses: threshold number of studies based on study sample size. *Evidence Based Medicine, 23*(Suppl 1): A21. https://doi.org/10.1136/bmjebm-2018-111024.21

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*, *339*. https://doi.org/10.1136/bmj.b2535

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Prisma-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*, 1-9. https://doi.org/10.1186/2046-4053-4-1

Moreira, M., Arcas, B., Sánchez, T., García, R., & Melero, M. (2023). Teachers' pedagogical competences in higher education: A systematic literature review. *Journal of University Teaching and Learning Practice*, *20*(1), 90-123. https://doi.org/10.53761/1.20.01.07

Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ, 372*. https://doi.org/10.1136/bmj.n71

Purvis A. Nicholas, V., & Tai, J. (2024). What's your problem? Writing effective research questions for quality publications. *Journal of University Teaching and Learning Practice. 21*(10). https://doi.org/10.53761/j64xa573

Roberts, K., Dowell, A., & Nie, J. (2019). Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC medical research methodology*, *19*(1), 1-8. https://doi.org/10.1186/s12874-019-0707-y

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*, 1-13. https://doi.org/10.1186/s13643-016-0315-4

Shepherd, D., & Suddaby, R. (2017). Theory building: A review and integration. *Journal of Management*, *43*(1), 59-86. https://doi.org/10.1177/0149206316647102

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*, 1-9. https://doi.org/10.1186/s12874-019-0782-0

Yang, D., Crawford, J., Daugaard, D. & Jia, J. (2025). Environmentally Sustainable Leadership: A Systematic Literature Review and Future Research Agenda. *Australian Journal of Management,* Advanced Online Publication.