NPS-PM-26-267


ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

# Evaluating Managerial Implications in Research Papers with Generative Pre-Trained Transformers

December 2025

**1st Lt. Louis V. Gianneschi, USN**

Thesis Advisors: Lt Col Jamie Porchia, Assistant Professor
Brett M. Schwartz, Lecturer

Department of Acquisition, Finance and Manpower

**Naval Postgraduate School**



Prepared for the Naval Postgraduate School, Monterey, CA 93943.




ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF ACQUISITION, FINANCE AND MANPPOWER
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Department of Acquisition, Finance and Manpower Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact the Acquisition Research Program (ARP) via email, arp@nps.edu or at 831-656-3793.


ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF ACQUISITION, FINANCE AND MANPPOWER
NAVAL POSTGRADUATE SCHOOL

# ABSTRACT

Academic management research has been criticized for prioritizing theoretical contributions rather than practical relevance for managers, creating a theory–practice gap. This exploratory study examined the strengths and weaknesses of utilizing artificial intelligence (AI) systems to evaluate managerial relevance. Ten papers published between 2015 and 2025 in leading supply chain and management journals were evaluated by six AI systems (ChatGPT-4o, ChatGPT-4o Deep Research, Grok-Fast, Grok-Expert, Opus 4.5, and Opus 4.5 Extended Thinking) across six criteria: (1) actionability, (2) novelty, (3) feasibility (problem-solving), (4) feasibility (resources), (5) impact, and (6) accessibility. Results showed that mean total scores varied widely across AI systems, ranging from 23.0 to 38.0, a difference of 15 points. Although all systems used the same evaluation rubric, they showed differences in scoring patterns, suggesting system-level biases related to model architecture. Papers from supply chain journals generally received higher scores, likely due to better alignment with the evaluation rubric. Overall, the findings suggest that current AI systems still struggle to apply complex, subjective criteria when assessing managerial relevance. As a result, hybrid approaches combining AI with expert human judgment are recommended for future applications in research evaluation.

THIS PAGE INTENTIONALLY LEFT BLANK

NPS-PM-26-267


ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

# Evaluating Managerial Implications in Research Papers with Generative Pre-Trained Transformers

December 2025

**1st Lt. Louis V. Gianneschi, USN**

Thesis Advisors: Lt Col Jamie Porchia, Assistant Professor
Brett M. Schwartz, Lecturer

Department of Acquisition, Finance and Manpower

**Naval Postgraduate School**



Prepared for the Naval Postgraduate School, Monterey, CA 93943.




ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF ACQUISITION, FINANCE AND MANPPOWER
NAVAL POSTGRADUATE SCHOOL

THIS PAGE INTENTIONALLY LEFT BLANK

# TABLE OF CONTENTS

# LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| AI | artificial intelligence |
| BERT | Bidirectional Encoder Representations from Transformers |
| DR | Deep Research |
| ET | Extended Thinking |
| GPT | generative pretrained transformer |
| ICC | intraclass correlation coefficients |
| JBL | *Journal of Business Logistics* |
| JOM | *Journal of Management* |
| JPSM | *Journal of Purchasing and Supply Management* |
| LLM | large language model |
| QWK | quadratic weighted kappa |

THIS PAGE INTENTIONALLY LEFT BLANK

# I. INTRODUCTION

The idiom "ivory tower" refers to a place where one is removed from real-world problems. Most often, it describes academics who focus on theory but offer little help to people outside their sphere of influence (Hawkins et al., 2022). According to Kieser et al. (2015), in management research, this criticism is common. While scholars debate models and publish in top journals, many managers rely on consultants or business articles to solve their problems. This creates a gap between what scholars produce and what practitioners need.

To these points, the concept of managerial relevance has lacked a clear definition. Jaworski (2011) stated that scholars have debated the theory–practice gap without defining what "managerial relevance" means. Debaters assume that the definition is obvious, leading to fragmented thinking. Without agreed-upon terminology, scholars approach the problem from different angles, creating inconsistent frameworks and misaligned priorities (Nicolai & Seidl, 2010). Jaworski (2011) emphasized that having a shared definition is vital for meaningful discussions between academics and practitioners.

Without a shared meaning, miscommunication occurs and both parties fail to engage in productive conversation and knowledge sharing. Clarifying this definition is the first step towards closing the theory–practice divide. For the sake of this thesis, *managerial relevance* is defined as the extent to which academic research can influence and improve decision-making and practice for those in managerial roles in an organization (Schauerte et al., 2023). This definition plays a key role in bridging the gap between theory and practice. In a perfect world, the role of business researchers is to produce knowledge that not only advances theory but also improves "real life" decisions. Managerial relevance must capture how scholarly findings connect with real-world management problems. Without relevance, even rigorous theoretical work may be ignored by practitioners, limiting the impact of research on business innovation and outcomes (Kieser & Leiner, 2009).

Academic research should not only advance theory but also impact practice. If research remains isolated in journals with little actionable advice, a relevance gap emerges. This gap has been a concern in both management and marketing fields. From a practitioner's perspective, academic studies are often seen as too abstract or a jungle of jargon (Kieser et al., 2015). Ankers and Brennan (2002) found that marketing managers knew very little about current academic research and felt that scholars did not understand the realities of business life. As a result, practitioners tended to rely on consultants or trade publications for insights, believing those sources better understood business realities and day-to-day operations.

This evidence shows a major disconnect: even when academic research has relevant findings, these findings may not reach nor resonate with managers. Therefore, academics must overcome communication barriers and focus on issues that matter to practitioners if they want their work to be used in everyday problems. From an academic perspective, the relevance gap has a different challenge. Scholars are rewarded for theoretical contribution and rigor, which can sometimes conflict with producing usable insights. As a result, there is an ongoing debate about how to balance rigor and relevance. Skeptics worry that catering to managerial interests might weaken scholarly work or lead to simplistic studies (Carton & Mouricou, 2017). Varadarajan (2020) addressed this concern by noting that relevance and rigor should not be viewed as a zero-sum trade-off. Sacrificing too much rigor for relevance can make research findings less valuable for all parties. Additionally, if a study does not have sufficient rigor, its recommendations to managers may be flawed. At the same time, Varadarajan (2020) emphasized that managerial relevance is an essential part of impact, and scholars must come to terms with this.

There are several challenges that make it difficult to evaluate managerial implications in research. First, there is the absence of a unified definition and framework for managerial relevance. As noted above, when scholars have different definitions of "relevance," it leads to inconsistency. Second, institutional reward systems prioritize theoretical contributions over practical applications, thus leading to misaligned incentives, which create challenges for evaluating relevance (Kerr, 1975). Third, the impact of research on practice is hard to measure. Fourth, while citation counts are tracked in

academia, there is no simple metric for how much a piece of research impacts outcomes. This makes it difficult to encourage truly impactful research. Fifth, there are organizational and communication barriers. Academics do not have strong incentives to invest time in simplifying their research for practitioners or engaging with managers to understand their problems. Conversely, managers may not have the time to skim through academic journals looking for their next solution.

## A. PROBLEM

In management and supply chain research, scholarly work aims to provide the field with both theoretical and practical contributions. However, a persistent gap exists between academic research and managerial practice. Research papers published in top-tier journals such as *International Journal of Logistics Management*, *International Journal of Physical Distribution and Logistics Management*, *Journal of Business Logistics*, *Transportation Journal*, *Journal of Marketing*, *Academy of Management Review*, and *Journal of Business Research* claim managerial relevance, yet they often lack operational guidance for managers (Hawkins et al., 2022). These managers have limited access to research that can truly be applied in day-to-day activities (Schauerte et al., 2023). Furthermore, there is a gap in how the quality of managerial implications is evaluated.

Regarding evidence to the problem, scholars such as Schauerte et al. (2023) have shown that published papers fail to propose useful insights for managers because they often focus on theories and lack sufficient real-world application. When research lacks clarity and feasibility, it is unlikely to be adopted in decision-making. Beyond the concern with practicality, there is also a knowledge gap in how managerial implications are evaluated across the field. Few studies attempt to score implications across multiple criteria. Moreover, existing evaluations tend to be manually performed and lack scalability. Emerging large language models (LLMs) such as generative pretrained transformer (GPT) have not yet been explored for their potential to support consistent and rapid evaluations of managerial relevance in academic contributions (Schauerte et al., 2023). Therefore, there is an opportunity to explore how GPT can serve as a tool to bridge the knowledge–practice gap more efficiently.

This thesis aims to make two contributions to this discussion. First, it tests whether a refined GPT model can serve as a viable tool for assessing the quality of managerial implications in management and marketing journals. Second, it tests how closely the GPT model is to human assessment of managerial implications. Together, these contributions offer value to scholars, managers, and even journal editors by improving the consistency and practicality of academic research.

## B. PURPOSE

This thesis aims to answer the following questions:

1. How well do recent research papers in top management and supply chain–related journals deliver strong managerial implications based on the criteria of Actionability, Novelty, Impact, Accessibility, Feasibility–Problem-Solving, and Feasibility–Not Resource Constrained?
2. What challenges exist when using a refined GPT model to score managerial implications in academic research?

## C. SCOPE

This study examines recent publications from 2015 to 2025 drawn from top-tier marketing and management journals. The articles were randomly selected from the *Journal of Management* (JOM), *Journal of Purchasing and Supply Management* (JPSM), and the *Journal of Business Logistics* (JBL) to ensure quality and impact.

## D. STUDY BENEFITS

This research addresses a critical gap in academic publishing by developing a systematic method to evaluate the practical relevance of management research. The need for this work stems from three interconnected problems: the theory–practice divide documented extensively in literature (Kieser et al., 2015; Varadarajan, 2020), the lack of standardized criteria for assessing managerial implications across journals (Schauerte et al., 2023), and the absence of scalable evaluation methods that can keep up with the volume of academic publishing.

Artificial intelligence (AI) is desired for this task because manual evaluation of managerial implications on a large scale is time-consuming. Journal editors and reviewers already face significant workload pressures, making comprehensive assessment of practical relevance across large numbers of submissions impractical (Hosseini & Horbach, 2023).

LLMs offer the potential for consistent, rapid assessment that can identify papers requiring deeper human expert review. Furthermore, AI systems can process and compare vast numbers of articles to identify patterns in how different fields, journals, or authors communicate practical applications (Sarker, 2022). These insights would be extremely complex for human reviewers to detect across hundreds of papers.

For academic researchers, this study provides concrete benefits. The adapted rubric testing offers scholars specific, actionable criteria for strengthening their managerial implications sections before submission. Rather than relying on reviewer feedback about "lack of practical relevance," researchers will understand exactly which dimensions require development (Kieser et al., 2015; Setkute & Dibb, 2025). Academics will better understand whether their implications lack sufficient detail, fail to provide actionable guidance, do not address feasibility, or suffer from unclear communication. This enables targeted improvement rather than vague rewriting.

For journal editors and reviewers, the study has multiple advantages. First, it provides standardized evaluation criteria that can be incorporated into reviewer guidelines and author instructions, improving consistency in how managerial relevance is assessed across submissions. Second, the AI-assisted assessment tool, once validated, serves as a screening mechanism that flags papers with weak managerial implications, allowing editors to provide early feedback to submissions that fail to meet the rubric's standards. Finally, by documenting current practices across top journals, the research shows which evaluation approaches work most effectively, enabling cross-journal improvement.

For practitioners and managers, improved managerial implications translate directly into better research usefulness. When academic papers provide clear, detailed, actionable guidance grounded in research, managers gain evidence-based solutions for

organizational challenges. This addresses the problem identified by Ankers and Brennan (2002), where practitioners abandon academic research in favor of consultants because scholarly work seems disconnected from business realities. By raising the bar for how research communicates practical applications, this work systematically ensures that management research generates value for organizations.

Finally, the broader academic community benefits from evidence about AI's capabilities and limitations in assessing complex dimensions like managerial relevance. As LLMs become increasingly integrated into scholarly systems, understanding where they excel and where they require human oversight becomes vital (Fuller & Bixby, 2024; Thelwall, 2024). This research contributes to responsible AI adoption in academic contexts by documenting specific reliability and validity considerations that may guide future research and applications.

## E. SUMMARY

To summarize, this chapter introduced the capstone project and outlined the main problem and purpose. It clarified the study's scope. This section also explained the expected benefits from this project's findings. Overall, this chapter provides the foundation for the chapters that follow.

# II. LITERATURE REVIEW

The integration of AI into academic research evaluation has the potential to be transformative for bridging the gap between management academia and day-to-day practice. While management research has dealt with the challenge of demonstrating practical relevance to business decision-makers, emerging AI technologies offer new possibilities for evaluating the managerial implications of academic work. LLMs and rubric-based assessment systems present opportunities that warrant investigation. The literature review that follows examines whether and how AI systems can effectively assess managerial relevance. This review synthesizes evidence from management science, educational technology, and computer science to evaluate both the potential and the limitations of AI relevance assessment. Furthermore, the review shows that while AI demonstrates capability in processing and classifying academic pieces with moderate reliability in structured environments, challenges regarding consistency, validity, and the assessment of concepts remain inconclusive.

## A. MANAGERIAL RELEVANCE DEFINITIONS AND DIMENSIONS

The concept of managerial relevance is a contested space in business scholarship. Schauerte et al. (2023) defined *managerial relevance* as "a research project's potential to influence managerial decision-making and thinking" (p.1). This definition emphasizes the practicality of academic work. The definition builds on Jaworski's (2011) concept of "role-relevant" research. Jaworski argued that scholarship should be crafted specifically to fit the needs of managerial roles rather than pursuing generic practicality. Through qualitative analysis of 65 in-depth interviews with senior editors, marketing managers, and academic researchers, Schauerte et al. (2023) identified four archetypes of relevance that can emerge from successful research. These include problem-solving relevance that directly addresses specific managerial challenges, explicating relevance that helps managers understand complex ideas, consolidating relevance that synthesizes existing knowledge into action, and forward-thinking relevance that anticipates future business challenges.

Nicolai and Seidl (2010) complement Jaworski's (2011) perspective by providing classifications of different relevance forms. They note that "although the notion of 'relevance' is frequently mentioned in literature, it is hardly ever defined and may have different, even contradictory, meanings in different contexts" (p. 1257). Their textual analysis of 450 articles in leading management journals revealed that relevance spans many dimensions. These include instrumental relevance (providing tools for action), conceptual relevance (offering new ways of thinking), and legitimative relevance (justifying existing practices; p. 1266).

These definitions and frameworks are not only discussed in theory but are beginning to be used in real-world settings. For example, the *Financial Times* recently introduced a "Research Impact" ranking that assesses how business school research contributes to policy, business, and society. The ranking incorporates nonacademic indicators such as mentions in patents, citations in policy documents, and references in the media. This signals a shift in the field toward valuing research that demonstrates practical influence and relevance (Jack & Dalal, 2025).

## B. THEORY–PRACTICE GAP

Furthermore, the foundations of managerial relevance are linked to the theory–practice gap (Kieser et al., 2015). Kieser et al. (2015) provided a 91-page review identifying two distinct paths of relevance literature: (1) studies that propose specific approaches for enhancing practical relevance and (2) literature examining the complex dynamic between management research and practice. Their analysis revealed that academic articles may contain useful information for practitioners yet remain highly unapproachable due to highly technical remarks, advanced statistical methods, and theoretical language extremely disconnected from business realities. This communication barrier thus reveals a deeper divide. Kieser and Leiner (2009) adopted a systems theory perspective to argue that the theory–practice gap may be "unbridgeable" because science and business organizations are separate social systems that cannot communicate across their boundaries. Each system operates according to its own logic, criteria, and codes, making genuine communication between academic and practitioner domains impossible.

Recent research has begun exploring how institutional structures and incentive systems influence the theory–practice divide. Arteaga et al. (2024) conducted a literature review identifying multiple complex challenges in bridging theory and practice. Their framework identifies diverse drivers. These drivers are motivation and values, knowledge accessibility, capacity constraints, language barriers, and funding limitations that shape the persistent divide between academic research production and practitioner knowledge utilization (Arteaga et al., 2024). They argued that closing the gap requires "integrating perspectives from multiple disciplines" (Arteaga et al., 2024, para. 2). These solutions include improved knowledge creation and accessibility, enhanced collaboration and communication between stakeholders, development of skills and capacity, contextual changes in organizational circumstances, and introduction of new boundary-spanning roles and tools (Arteaga et al., 2024). This framework and perspective suggest that no single driver can fully resolve the fundamental tensions between academic and practitioner communities.

## C. PRACTITIONERS' CONSIDERATIONS

This view contrasts with scholars advocating for a more engaged approach to bridge the divide (Bansal et al., 2012; Bartunek & Rynes, 2014). For example, Van de Ven (2007) proposed engaged scholarship as "a participative form of research for obtaining the different perspectives of key stakeholders [...] in studying complex problems" (p. 9). He argued that academics can achieve both the goals of rigor and relevance through collaboration with practitioners throughout the research process. His framework emphasizes four fundamental activities: problem formulation, theory building, research design, and problem-solving. All these activities must be conducted in partnership with practitioners to ensure research addresses genuine business challenges while maintaining the academic soundness that scholars desire.

The interpretation of managerial relevance in marketing, management, and supply chain research highlights key characteristics that differentiate it from academic contribution. First, Kieser et al. (2015) noted that "because practitioners do not understand the mathematics and statistics that characterize most contemporary research, many of the

articles published in academic research journals today might as well be written in Greek" (p. 153) Second, relevance requires actionability. Research should provide clear implications for managerial decision-making rather than purely descriptive findings or abstract propositions (Kieser et al., 2015). Third, relevant research needs to have timeliness. It should address current or emerging business challenges rather than solely pursuing interests primarily for academic audiences (Kieser et al., 2015). Fourth, truly relevant research has awareness. Managerial effectiveness depends on organizational and environmental situations that theory must accommodate (Kieser et al., 2015). Ankers and Brennan's (2002) study of practitioner perspectives revealed the large disconnect between academics' desires and practitioners' perceptions. Their interviews with experienced marketing practitioners found that business managers "knew very little about the current state of academic research in marketing and considered that academic researchers did not understand the realities of business life and could not communicate effectively with managers" (Ankers & Brennan, 2002, p. 15). Furthermore, practitioners expressed clear preference for working with consultants over academics, viewing consultants as more grounded to business realities and more effective as communicators (Ankers & Brennan, 2002). This study revealed that the relevance gap encompasses not only research content but also issues of communication, trust, and perceived legitimacy of knowledge in business settings.

## D. WHY MANAGERIAL RELEVANCE MATTERS

The significance of managerial relevance extends beyond academic debates. It carries consequences for research utilization, decision-making, and societal impact. Nobel (2016) explored the long-standing divide between academic research and business practice by interviewing leading scholars and business executives. The purpose of Nobel's article was to understand why much of management research fails to influence managerial decision-making or address real-world challenges.

Among those interviewed, Michael Toffel, the Senator John Heinz Professor of Environmental Management and faculty chair of the Harvard Business School Business and Environment Initiative, stated, "This is my soapbox message to academics: be more

relevant" (Nobel, 2016, para. 4). He argued that while most business scholars recognize primary duties of teaching students and generating new knowledge through research, "the lack of practical relevance of much of our research might suggest that few of us also have the ambition to improve the decisions of the managers and policymakers whose actions we study" (Nobel, 2016, para. 5). This statement suggests that research has limited influence on actual management practices.

The issue extends beyond the organizational level. According to Nobel (2016), society loses valuable insight that could help address complex challenges when research fails to reach practitioners. Journalist Nicholas Kristof criticized academia for producing thinkers who "just don't matter in today's great debates" (Nobel, 2016, para. 7). Similarly, Professor Andrew Hoffman suggested that many academics do not view their role as "an enabler of direct public participation in decision-making" (Nobel, 2016, para. 8). This disconnection at the societal level could lead to questions about the true value and purpose of academic research.

From the practitioner perspective, the relevance gap creates significant barriers to evidence-based decision-making. Business leaders are frustrated with academic research accessibility and usability. Donovan Neale-May, executive director of the Chief Marketing Officer Council, stated that academic research "tends to be overly complex, hard to digest, and not backed by real quantitative insights" (Nobel, 2016, para. 2). These practitioners' statements reveal that for many business leaders, the core issue is not the lack of research but the lack of usable and relevant research.

According to Nobel (2016), this gap can be narrowed when scholars collaborate with practitioners. Toffel argued that engaging practitioners early "not only helps improve the research but also increases the likelihood that practitioners will subsequently read and appreciate a translation of that work" (Nobel, 2016, para 20). Harvard Business School scholars reinforce this idea. Francesca Gino noted that many of her most impactful projects were inspired by real-world challenges, while Benjamin Edelman admitted that writing only for academic audiences does not motivate him (Nobel, 2016). Collectively, these perspectives show that research can be both rigorous and relevant, but meaningful engagement with practice is essential.

From the perspective of journal editors and the academic publishing ecosystem, managerial relevance must have a role in shaping editorial priorities and journal entries. Leading journals have implemented specific mechanisms to enhance relevance. These include requirements for managerial abstracts that translate findings for practitioner audiences, encouragement of video summaries to nonspecialists, and dedicated sections for practically oriented work. *Academy of Management Perspectives* requires that published papers "inform an issue of evident importance to managerial practice and/or policy" while engaging "in rigorous and original conceptual or empirical analysis" that "concisely and clearly convey key ideas to a non-specialized audience," emphasizing that research must be "relevant, rigorous, and readable" (Academy of Management, n.d.). The journal *Supply Chain Management: An International Journal* emphasizes the dual importance of academic rigor and practical impact, stating that "research findings must demonstrate international relevance and global impact to both theory and practice" (Emerald Publishing, n.d., "Scope" section). As Horbach and Halffman (2020) noted, editorial practices are shaped by not only academic rigor and relevance but also by the commercial considerations that incentivize impact within publishing ecosystems. These findings suggest that relevance emerges from collaborative processes that engage academic and practitioner communities throughout the research life cycle.

## E. ARTIFICIAL INTELLIGENCE JUDGEMENT LIMITATIONS

The rapid advancement of AI technologies has generated a wave of interest in leveraging AI systems in academic research (Cao et al., in press). LLMs like GPT-4, Claude, Grok, and Bidirectional Encoder Representations from Transformers (BERT) present particular interest (Lin, 2025). Current applications of AI in academic content evaluation span multiple academic research and publication processes. These include peer review assistance, abstract screening, research quality assessment, and citation analysis (Ge et al., 2024; Liang et al., 2023). However, there is limited discussion regarding AI's capability to assess managerial implications. The following literature shows that while AI technologies are effective at classifying research content, they remain limited in assessing its practical applicability.

Checco et al. (2021) conducted foundational research on AI-assisted peer review by training neural networks on 3,300 conference papers to predict peer review outcomes. Their models achieved successful prediction of acceptance or rejection decisions using metrics. These metrics included word frequencies, readability scores, and formatting characteristics. The strong correlation between document features and review outcomes demonstrates that AI can approximate human evaluation patterns. However, the models approximate human decisions using heuristics rather than evaluating scientific merit. The authors noted that their approach "raises concerns about perpetuating biases" (Checco et al., 2021, p.3). AI systems trained on historical peer review decisions will replicate whatever systematic biases human reviewers exhibited in the training data. These biases may relate to writing style, institutional prestige, or preferences.

## F. GENERAL CONSIDERATIONS IN ARTIFICIAL INTELLIGENCE USAGE

The following section reviews findings from several investigations that tested AI systems on research evaluation tasks. To begin, Thelwall (2024) tested ChatGPT-4 on 51 articles using UK Research Excellence Framework 2021 scoring criteria. Regarding insights, the research found that individual scores showed only weak correlation ($r = 0.281$) with expert self-evaluations. However, averaging scores over 15 iterations improved correlation substantially ($r = 0.509$). This suggests that LLM variability can be partially mitigated through aggregation. The significance is that "ChatGPT struggles to make fine-grained evaluations" (Thelwall, 2024, p.1). The model cannot effectively distinguish between moderately good and excellent research. The model can extract claims about rigor and originality from papers, but it lacks the expertise to evaluate whether those claims are justified. Thelwall (2024) concluded that ChatGPT "does not yet seem to be accurate enough to be trusted for any formal or informal research quality evaluation tasks" (Thelwall, 2024, p.1). This limitation is particularly relevant for assessing managerial relevance.

Furthermore, a Stanford study examined whether AI systems like GPT-4 could provide useful feedback on research papers. The researchers found that approximately 31% of GPT-4's comments aligned with those of human reviewers. To verify the AI was

analyzing the content rather than generating random feedback, researchers tested it with shuffled papers. Accuracy dropped from 58% to just 1%, confirming the system was responding to actual paper content (Liang et al., 2023).

However, the feedback had significant limitations. Comments were often "too vague," lacking "specific technical areas of improvement" and showing insufficient "depth on architecture and design" (Phys.org, 2023, para. 15). This indicates that while AI models can identify relevant aspects of research papers, they cannot yet provide detailed, actionable guidance that expert human reviewers offer. The technology shows promise for preliminary screening but falls short of replacing expert review. The researchers concluded that "expert human feedback will still be the cornerstone of rigorous scientific evaluation," (Liang et al., 2023, p. 7).

Machine learning shows both promise and clear limits when predicting research quality. Thelwall (2023) tested AI on thousands of articles from the UK's Research Excellence Framework, which had already been scored by expert reviewers. The AI predicted article quality with up to 72% accuracy, though results varied by field. Medical sciences, physical sciences, and economics performed best, while arts, humanities, and mathematics showed lower accuracy. The model analyzed article metadata, citations, author backgrounds, and abstracts.

However, Thelwall (2023) noted, "Substantially higher levels of accuracy may not be possible due to the need for tacit knowledge to understand the context of articles to properly evaluate their contributions" (p. 568). In other words, AI lacks the specialized expertise and contextual understanding that human reviewers possess. This limitation is particularly important for assessing managerial relevance.

Williams and Grant (2023) tested AI's ability to evaluate research impact using multiple models, achieving up to 90.4% accuracy when combining different features. However, they identified a critical problem: the models perpetuated existing biases rather than making objective assessments. For evaluating managerial relevance, this means AI systems may simply favor already highly rated research instead of genuinely assessing

practical value. These biases can include preferential treatment of prestigious authors, well-funded institutions, or writing styles.

Mostafapour et al. (2024) examined GPT-4's potential to "assist in scholarly tasks, including conducting literature reviews" (p. 1). In a review of the Mostafapour et al. study, Evans (2025) summarized the findings, noting that "the results showed that GPT-4 generated a broad list of 21 relational factors in seconds, identifying a wide breadth of content and responding rapidly. The human review, while much slower, provided deeper contextualization, greater accuracy, and clearer methodological transparency" (para. 4). Evans (2025) concluded that GPT-4 can support but "cannot replace expert human scholarship" (para. 4).

Additional literature reveals gaps in AI's capability of assessment, often needing context. AI struggles with several key areas such as tacit knowledge and originality (Thelwall, 2023). Collectively, the literature supports the need for a hybrid teaming approach between AI and human experts to extract meaningful insights. As such, this research combines AI in initial screening with human expert judgment to determine relevance, originality, and practical applicability.

## G. EVOLUTION OF RUBRIC-BASED ASSESSMENT IN ARTIFICIAL INTELLIGENCE SYSTEMS

The history of rubric-based AI assessment spans over 60 years. It has gone from simple rule-based systems to neural networks and LLMs. Challenges that remain include performance, accuracy, and standardization. These challenges are relevant for research evaluation (Hussein et al., 2019).

One of the earliest automated essay scoring systems was Project Essay Grade (PEG), developed by Ellis Batten Page in 1966 (Hussein et al., 2019). PEG used a rule-based approach that analyzed patterns to predict essay quality. The system first trained on prescored essays to generate statistical weights, then applied these weights to score new essays. Although PEG achieved a 0.87 correlation with human raters, the limited computing technology of the 1960s made it too expensive to implement widely (Hussein et al., 2019). This cost barrier halted development for approximately 20 years (Hussein et

al., 2019). Despite this setback, PEG established foundational principles still used in modern AI systems, including training on human-scored examples, identifying statistical patterns, and applying learned patterns to evaluate new content (Hussein et al., 2019).

The modern era of automated essay scoring began in the late 1990s with three major systems. The Intelligent Essay Assessor (IEA), developed by Landauer, Foltz, and Laham and released in 1997 (Hussein et al., 2019), required only about 100 prescored training essays compared to 300 to 500 for competing systems and achieved a 0.90 correlation with human raters in a middle school assessment of 800 students (Hussein et al., 2019).

Educational Testing Service developed E-rater in 1998. E-rater combined statistical analysis with natural language processing using an 11-feature rubric structure (Hussein et al., 2019). The system trained on human-scored essays before evaluating new submissions, achieving correlations between 0.87 and 0.94 with human assessors. This reliability was sufficient for high-stakes standardized tests like the graduate record examination (GRE; Hussein et al., 2019).

Vantage Learning's IntelliMetric, released in 1998, was the first automated essay scoring system to explicitly use AI to simulate human scoring. IntelliMetric demonstrated impressive multilingual capability, scoring essays in 10 languages. With an average correlation of 0.83 with human raters, IntelliMetric proved that AI-based approaches could match traditional statistical methods (Hussein et al., 2019).

Between 2016 and 2018, neural network approaches advanced automated essay scoring significantly. Alikaniotis et al. (2016) developed a model, achieving strong correlations of 0.91 (Spearman) and 0.96 (Pearson) that surpassed earlier systems. Taghipour and Ng (2016) integrated multiple processing layers, achieving a 5.6% improvement over baseline systems (Hussein et al., 2019). Dasgupta et al. (2018) introduced a model incorporating linguistic, psychological, and cognitive features, achieving a Pearson correlation of 0.94, Spearman correlation of 0.97, Cohen's $\kappa$ of 0.97, and a root mean square error of 2.09.

The emergence of LLMs in the 2020s represents the latest phase in automated assessment. Liew and Tan (2024) explored GPT-based automated essay grading and found

that GPT-3.5 Turbo, without any specialized training, achieved a Quadratic Weighted Kappa (QWK) of 0.68 for essay scoring. Performance depended heavily on prompt engineering, suggesting that LLMs may learn quality criteria during their initial training, potentially reducing the need for explicit rubrics in some applications.

This historical progression reveals a fundamental trade-off between objectivity and performance of the models. Earlier approaches to automated essay grading that relied on manually designed features and rules created by human experts have closer alignment with explicit rubric criteria (Hussein et al., 2019). They maintain fairness and transparency but require substantial manual effort. Automatic learning systems achieve higher predictive accuracy but may have subjective factors of human biases. This makes it challenging to verify that assessments genuinely reflect rubric dimensions (Hussein et al., 2019). For research evaluation uses, this trade-off suggests that achieving both high performance and meaningful transparency about what constitutes "relevance" remains a challenge.

## H. CONSISTENCY AND ACCURACY CHALLENGES IN ARTIFICIAL INTELLIGENCE RUBRIC-BASED ASSESSMENTS

Despite advances in AI assessment capabilities, evidence documents persistent challenges with consistency across models and iterations, accuracy in scoring, and multiple forms of systematic bias. These limitations have direct implications for applications in research evaluation, specifically for assessing managerial relevance.

Fuller and Bixby (2024) showed inconsistency in LLM grading systems. Testing ChatGPT and Claude AI on identical assignments uploaded multiple times revealed concerning results. ChatGPT produced a 24-point difference between lowest and highest scores (74% to 98%) on the same assignment. Claude AI showed a 33-point difference (58% to 91%). Both systems produced variations not only in numerical scores but also in feedback rationale for identical rubric items. ChatGPT rated reference pages as "exemplary" when no reference page existed in the assignment. The authors noted that these systems are "designed for creativity, not assessment accuracy, leading to hallucinations and false affirmatives" (Fuller & Bixby, 2024, para. 8). Extreme score variations undermine reliability. Different results from the same model on repeated queries

compound reliability problems. Poličar et al. (2025) noted that "due to their probabilistic nature, LLMs can generate different grading responses when queried multiple times" (p. 10).

Zhang et al. (2024) examined LLM performance for criterion-based grading from agreement to consistency. The study found that the most optimized prompt achieved only moderate absolute agreement (ICC = 0.68 to 0.74) with official benchmarks. LLMs provided significantly lower grades than official benchmarks. The authors concluded that domain-specific knowledge fundamentally limits precision, not model complexity. This finding suggests that general-purpose LLMs may lack the specialized expertise required to assess dimensions like managerial relevance.

Levman et al. (2023) revealed a significant finding regarding error consistency research. The authors stated, "Even in situations where all models created as part of validation yield the same (or almost the same) accuracies, it is possible for the various trained validation models to largely disagree on which samples they make their mistakes on" (Levman et al., 2023, para. 3). This means that different AI model iterations may make errors on research papers even when overall accuracy appears identical. For research evaluation, this shows that multiple researchers with equivalent work quality might receive different assessments depending on which model version evaluates their work. This is a concern.

Chai et al. (2024) documented that while students perceive AI as fairer due to apparent objectivity and transparency, "subjective bias of the evaluator might still be hidden in the algorithm" (Chai et al., 2024, p. 2) In summary, AI rubric systems exhibit multiple bias forms. These include historical biases where training data reflects past human biases, representation bias from training data, and measurement bias.

The limitations documented in the literature review emphasize context and domain constraints. As a result, the literature reviewed consistently recommends hybrid approaches. These approaches include using AI for tracking, implementing human oversight and validation, providing detailed criteria, and monitoring for biases. For evaluating managerial relevance specifically, these challenges suggest that current AI

systems lack the reliability, validity, and freedom from bias necessary for accurate assessment.

Daugherty and Wilson (2024) introduced the concept of reciprocal apprenticing as one of several "fusion skills" necessary for working effectively with generative AI in complex decision-making environments (p. 1). Rather than treating AI as just a tool, reciprocal apprenticing involves a dynamic process in which users teach AI systems through feedback and refinement while simultaneously learning how to better interact with those systems. This two-way relationship creates growth on both sides: the AI becomes more context-aware and better aligned with tasks and goals, while the human makes improvements in leveraging the AI's strengths without overreliance.

Applied to the problem of evaluating managerial relevance, reciprocal apprenticing offers a practical response to concerns about inconsistency, inaccuracy, and bias. Managerial relevance is nuanced and multidimensional. While AI demonstrates impressive capabilities in processing, classifying, and analyzing research content, there are still issues with consistency, accuracy, and bias. This suggests that current AI systems and technology are not capable of these tasks without constant human intervention. Reciprocal apprenticing allows that intervention to be iterative (Daugherty & Wilson, 2024).

THIS PAGE INTENTIONALLY LEFT BLANK

# III. METHODOLOGY

This chapter outlines the methodology used to conduct this study on evaluating managerial relevance through AI-based scoring. The selection of articles was based on 10 random samples from three top-tier journals—JBL, JPSM, and JOM—published between 2015 and 2025. This timeframe was chosen to ensure timely insights and relevance.

The adaptation of the rubric is noted, and the AI scoring process is detailed. Analytical methods include coefficient of variation and comparison of scoring patterns across evaluation dimensions Due to the exploratory nature of this research, this methodology provides a practical and adaptable framework for comparing emerging AI capabilities.

## A. RESEARCH DESIGN OVERVIEW

This study examined the ability of AI systems to evaluate managerial relevance in academic management research. The study tested three AI tools and their different extended thinking modes to determine which model produced the most consistent and valid results using a rubric framework adapted from Hawkins and Muir (2024). The evaluation criteria included six dimensions: Actionability, Novelty, Impact, Accessibility, Feasibility (Problem-Solving), and Feasibility (Not Resource Constrained). Each dimension was scored on a 1 to 7 Likert scale, with total scores calculated as the sum of individual dimension scores (maximum possible total of 42 points).

## B. ARTIFICAL INTELLIGENCE SYSTEM SELECTION

Three AI systems and their different extended thinking were selected. ChatGPT4o represented OpenAI's standard GPT-4o model. ChatGPT4o DR employed OpenAI's deep research capability for more thorough evaluation (OpenAI, 2025). Grok-Fast represented xAI's standard model. Grok-Expert utilized the same Grok architecture with expert-level thinking capabilities (xAI, n.d.). Opus 4.5 represented Anthropic's standard model. Opus 4.5 ET employed extended thinking capabilities for deeper analysis (Anthropic, 2025).

## C. DATA COLLECTION

A sample of 10 research articles from JBL, JPSM, and JOM, published between 2015 and 2025, were selected using simple random sampling with exclusion criteria (Thompson, 2012). A nearly equal split of four from JBL, three from JPSM, and three from JOM were chosen. The sampling process involved two sequential random selections: first, journal issues were randomly selected using a random number generator, and second, individual articles within those issues were randomly selected using a second random number. Editorials were excluded from the sample prior to selection, limiting the population to research articles only. This method ensured that each research article had an equal and known chance of selection (Thompson, 2012).

## D. ANALYTICAL APPROACH

Four analytical steps were utilized to examine AI evaluation reliability and scoring patterns. First, mean total scores for each paper were compared across all six AI systems, and the coefficient of variation (CV) was calculated as a standardized measure. Papers with CV values below 10% were classified as high agreement, 10–20% as moderate agreement, and above 20% as low agreement (Reed et al., 2002). Second, AI model scoring was examined for each of the dimensions defined by Hawkins and Muir (2024) to identify which dimensions showed the greatest agreements and disagreements across systems. Third, AI model mean scores were compared to determine which systems applied more generous or strict scoring. Fourth, each AI system's average scoring per dimension was analyzed to identify dimension-specific patterns.

## E. VALIDITY AND BIAS

Several potential biases should be brought to attention. The use of exclusion criteria limits external validity, as results can only be applied to research articles and not to all journal content (Pannucci & Wilkins, 2010). Additionally, if the distinction between research articles and editorials was inconsistent across the different journals and issues, this may have introduced selection bias into the process (Pannucci & Wilkins, 2010). Despite these challenges, random selection minimizes systematic bias and provides a statistically solid foundation (Thompson, 2012).

## F. SUMMARY

In summary, this study evaluated the ability of artificial intelligence systems to assess managerial relevance in academic management research. Ten research articles published between 2015 and 2025 (with editorials excluded) were selected through simple random sampling from three journals (JBL, JPSM, and JOM). Three AI systems and their different extended thinking models (ChatGPT4o, ChatGPT4o DR, Grok-Fast, Grok-Expert, Opus 4.5, and Opus 4.5 ET) evaluated these articles using an adapted rubric based on Hawkins and Muir (2024). The rubric assessed six criteria: Actionability, Novelty, Impact, Accessibility, Feasibility (Problem-Solving), and Feasibility (Not Resource Constrained). Analytical methods included coefficient of variation, comparison of AI model averages, and analysis of dimension-specific scoring patterns across systems. Chapter IV presents the findings of this exploratory study and addresses the research questions proposed in Chapter I.

THIS PAGE INTENTIONALLY LEFT BLANK

# IV. RESULTS

## A. FINDINGS

### 1. Research Question 1

The first research question examined how well recent research papers in top management and supply chain related journals deliver strong managerial implications based on the criteria of Actionability, Novelty, Impact, Accessibility, Feasibility (Problem-Solving), and Feasibility (Not Resource Constrained). To address this question, ten papers published between 2015 and 2025 in the *Journal of Management* (JOM), *Journal of Business Logistics* (JBL), *Journal of Purchasing and Supply Management* (JPSM) were randomly selected from an archive of 200 papers from these three journals. These ten papers were evaluated using six different artificial intelligence systems. Each AI evaluator assessed papers on a 1 to 7 Likert scale across six dimensions, with 1 representing strongly disagree and 7 representing strongly agree. The total scores were calculated as the sum of individual dimension scores (maximum possible total of 42 points).

The six AI evaluation systems used the same prompting strategy after the rubric adapted from Hawkins and Muir (2024) was attached:

"Act as Research Relevance Rater. Before responding, read the entire prompt. Using a fixed 1 to 7 agreement scale, evaluate a research paper across all rubric categories. Your role and identity are fixed. You are Research Relevance Rater, a GPT system that scores research papers using only retrieved text. You do not invent material. Accessibility is the only category where web browsing is allowed. Use the item lists and factors provided in the uploaded files. Your scoring must reflect the rubric exactly. Use direct and plain language in all outputs. Do not use words like delve, moreover, this paper aims, or our aim. Keep evidence notes short with one or two sentences. Provide:

1. A score from 1 to 7 for each rubric category.
2. A short evidence note for each score that cites retrieved text only.
3. A final summary that lists the scores in order.

This prompt was influenced by an email conversation from Dr. Daniel Finkenstadt and from the work of Koraishi (2024), who examined the reliability of ChatGPT in evaluating IELTS Writing Task 2 responses (Koraishi, 2024) (Dr. Finkenstadt, personal communication, September 8, 2025).

Regarding AI system selection, ChatGPT4o represented OpenAI's standard GPT-4 model. ChatGPT4o Deep Research (DR) utilized deep reasoning capabilities with more in-depth evaluation (OpenAI, 2025). Grok-Fast represented xAI's Grok model with standard speed. Grok-Expert used the expert-level evaluation (xAI, n.d.). Opus 4.5 represented Anthropic's Claude Opus model with standard configuration. Opus 4.5 Extended Thinking (ET) employed extended thinking capabilities for a more comprehensive analysis (Anthropic, 2025). The AI platforms chosen for this study were intentionally selected to represent a range of different system designs and capability levels currently available in the commercial AI market. As important, was the decision to evaluate both the base versions and enhanced reasoning versions of each AI platform. If enhanced reasoning consistently improved scoring reliability, it would suggest that using more advanced AI configurations provides real benefits for complex evaluation tasks. Overall, this multi-rater design allowed assessment of paper relevance across AI systems.

Table 1 presents the complete evaluation results from all six AI systems, organized to show the range of scores each paper received across different evaluators. Papers are listed in descending order by their mean total score across all six AI raters to allow easy identification of consistently high-performing and low-performing research.

Examination of mean total scores revealed variation in assessed managerial relevance across the ten-paper sample. Mean scores ranged from 26.3 to 35.5 out of 42 possible points, representing a 9.2-point range. The sample mean across all papers and all AI raters was 31.4 (SD = 3.78). This shows that the AI raters were able to distinguish papers with relatively higher managerial relevance from those with lower relevance. Nonetheless, differentiation across individual papers exists.

The standard deviation column in Table 1 provided critical information about inter-rater agreement. While there are no universally accepted cutoff values for standard

deviation on rating scales, prior measurement research suggests using the coefficient of variation (CV = SD/Mean × 100) as a standardized way to assess variability relative to the mean (Bland & Altman, 1996; Reed et al., 2002). In general, CV values below 15% are considered acceptable for subjective assessments, while higher values raise concerns about reliability (Bland & Altman, 1996; Reed et al., 2002).

The paper with the lowest variability was Guntuka et al. (2024), with a standard deviation of 3.06 and a CV of 9.6%, indicating strong agreement across the six AI systems. In contrast, Jain and Huang (2020) showed the highest variability, with a standard deviation of 6.83 and a CV of 26.0%. Papers with CV values below 12.7% (corresponding to standard deviations below 4.0, given the sample mean) generally received consistent scores, while papers with CV values above 15.9% (standard deviations above 5.0) showed disagreement.

The range column showed the spread between the most generous and most strict AI evaluator for each paper. Ranges spanned from 7 points (Essuman et al., 2023) to 20 points (Jain & Huang, 2020). These significant ranges indicated that AI evaluators disagreed about some papers' managerial relevance. This variation raised important questions about AI evaluation reliability and the sources of said disagreement across models.

Papers receiving consistently low scores (mean below 28 across all AI systems) included Jain & Huang (2020) and Mitchell et al. (2023). These papers scored poorly even from generous systems, suggesting weaknesses in managerial implications that all AI systems detected.

Papers exhibiting highly divergent scores across AI systems included Paruchuri et al. (2024), which received scores ranging from 20 (ChatGPT4o) to 36 (ChatGPT4o DR) and Pal & Altay (2019), which spanned from 26 (Opus 4.5 ET) to 40 (Grok-Expert). This wide variability suggested either that these papers contained ambiguous managerial implications or that AI systems interpreted evaluation criteria differently when assessing these specific papers.

Table 1. Distribution of Managerial Relevance Scores across Six Artificial Intelligence Evaluators

| Paper (Short Title) | Authors (Year) | Journal | Mean Total | SD Total | Min Total | Max Total | Range |
|---|---|---|---|---|---|---|---|
| Developing and deploying competences for innovative public procurement | Taheriruh et al. (2025) | JPSM | 35.5 | 4.55 | 28 | 41 | 13 |
| Low power, high ambitions: New ventures | Bjørgum et al. (2021–2022) | JPSM | 35.0 | 3.58 | 31 | 40 | 9 |
| In search of operational resilience | Essuman et al. (2023) | JBL | 34.7 | 3.20 | 32 | 39 | 7 |
| Implementing PBC as procurement strategy | Molitoris et al. (2025) | JPSM | 32.2 | 5.38 | 27 | 40 | 13 |
| Identifying Key Success Factors BoP | Pal and Altay (2019) | JBL | 32.0 | 5.10 | 26 | 40 | 14 |
| Supply Chain Characteristics influence rival responses | Guntuka et al. (2024) | JBL | 31.8 | 3.06 | 28 | 37 | 9 |
| Pack-and-a-Half Rule eliminate backroom inventories | Eroglu et al. (2018–2019) | JBL | 31.0 | 3.63 | 29 | 38 | 9 |
| Governance Failure and Governance Under Failure | Paruchuri et al. (2024) | JOM | 27.7 | 6.44 | 20 | 36 | 16 |
| Oh the Anxiety! Supervisor Bottom-Line Mentality | Mitchell et al. (2023) | JOM | 27.5 | 3.21 | 22 | 35 | 13 |
| Learning from the Past: Scientist Relocation | Jain and Huang (2020) | JOM | 26.3 | 6.83 | 17 | 37 | 20 |

*Note*. Mean, SD, Min, Max, and Range calculated across six AI evaluators (ChatGPT4o, ChatGPT4o DR, Grok-Fast, Grok-Expert, Opus 4.5, Opus 4.5 ET). Total scores represent sum of six dimension scores (Actionability, Novelty, Feasibility-Problem, Feasibility-Resources, Impact, Accessibility), with maximum possible score of 42.

Analysis of journal-level patterns revealed differences in assessed managerial relevance across outlets. Papers published in JPSM (n=3) achieved the highest mean total score of 34.2 across all AI raters. JBL papers (n=4) averaged 32.4, while JOM papers (n= 3) produced a substantially lower mean of 27.2. The 7.0-point gap between JPSM and JOM may have reflected how well the journal content matched the evaluation rubric, rather than differences in research quality. JOM's stronger focus on theory may have led to lower scores, not because the research was less valuable, but because the evaluation framework emphasized practical application. On the other hand, JPSM articles may have more closely aligned with the rubric and thus received higher scores. Furthermore, the lower average

scores and higher variability for JOM papers further emphasizes that the evaluation rubric may be better suited for practice-oriented research, and that assessing managerial relevance in theory-driven papers may possess additional challenges for AI systems.

Table 2 presents dimension-level statistics calculated across all papers and all AI raters, revealing which evaluation criteria showed strongest and weakest performance across the sample.

Table 2. Aggregate Statistics for Six Evaluation Dimensions across All Papers and Artificial Intelligence Raters

| Dimension | Mean | SD | Min Observed | Max Observed | Most Common Score |
|---|---|---|---|---|---|
| Actionability | 5.23 | 1.42 | 2 (ChatGPT4o) | 7 (Grok-Expert) | 5 |
| Novelty | 5.62 | 0.68 | 3 (Opus 4.5 ET) | 7 (ChatGPT4o DR) | 6 |
| Feasibility (Solve Problem) | 4.73 | 1.31 | 2 (ChatGPT4o) | 7 (Grok-Expert) | 5 |
| Feasibility (Not Resource Constrained) | 4.95 | 1.36 | 2 (ChatGPT4o) | 7 (Grok-Expert) | 5 |
| Impact | 5.45 | 1.09 | 3 (ChatGPT4o) | 7 (Grok-Expert) | 5 |
| Accessibility | 5.93 | 1.38 | 3 (ChatGPT4o) | 7 (Grok-Expert) | 7 |

*Note*. Statistics calculated across 60 total observations (10 papers × 6 AI raters) for each dimension.

Accessibility emerged as the highest-scoring dimension with a mean of 5.93, though with variability (SD = 1.38). The most common Accessibility score was 7, achieved in 43% of all evaluations (26 of 60 observations). This percentage shows that there are constraints that limit accessibility. As to why ChatGPT4o rated so low for Accessibility, it stated, "The paper received a 3 for Accessibility because its academic language and dense theoretical framing make it difficult for managers to extract clear, practical insights," (OpenAI, 2025).

Novelty demonstrated the second-highest mean (5.62) and the lowest variability (SD = 0.68) of any dimension. The restricted variance suggested that papers maintained relatively uniform standards for novelty.

Impact scores averaged 5.45 with moderate variability (SD = 1.09). The concentration of scores around 5 and 6 (74% of all observations) indicated that AI

evaluators generally perceived research as offering organizational benefits as described in the rubric.

Actionability produced a mean of 5.23 but with the highest variability (SD = 1.42) among dimensions. The wide range from minimum 2 to maximum 7 showed there were substantial differences depending on how each model interpreted actionable recommendations.

Both Feasibility dimensions showed lower means (4.73 and 4.95) compared to other criteria, indicating that feasibility was a weakness across the sample. AI evaluators appeared less confident that research offered feasible solutions. Similar standard deviations (1.31 and 1.36) suggested that both Feasibility dimensions varied by about the same amount across papers, whether the dimension focused on solving the problem or on resource constraints.

To examine how different AI systems evaluated the same papers, Table 3 presents mean scores for each AI rater across all ten papers, revealing systematic differences in evaluation scoring patterns.

Table 3. Mean Total Scores by Artificial Intelligence Evaluation System across All Ten Papers

| AI System | Mean Total Score | SD | Min Score | Max Score | Range |
|---|---|---|---|---|---|
| Grok-Expert | 38.0 | 2.21 | 34 | 41 | 7 |
| ChatGPT4o DR | 32.6 | 4.72 | 24 | 38 | 14 |
| Opus 4.5 | 31.6 | 3.57 | 25 | 37 | 12 |
| Grok-Fast | 30.8 | 2.57 | 27 | 35 | 8 |
| Opus 4.5 ET | 28.9 | 3.57 | 20 | 35 | 15 |
| ChatGPT4o | 23.0 | 4.22 | 17 | 32 | 15 |

*Note*. Statistics calculated across ten papers for each AI system. Systems ranked from most generous (highest mean) to most strict (lowest mean).

Grok-Expert emerged as the most generous evaluator with a mean total score of 38.0 across all papers, positioning on average the journal articles at 90% of maximum possible points. This system also demonstrated the narrowest range (7 points). The combination of high means and low variance indicated that Grok-Expert applied generous standards with limited discrimination. ChatGPT4o was the strictest evaluator, producing a

mean total score of only 23.0 (55% of maximum). However, ChatGPT-4o also showed one of the widest scoring ranges (15 points). Although this system applied stricter evaluation overall, it was still able to distinguish between papers. Due to the lack of clustering scores, it may have made better use of the entire rubric.

The systems in the middle (ChatGPT4o DR, Opus 4.5, Grok-Fast, Opus 4.5 ET) produced mean scores between 28.9 and 32.6. ChatGPT4o DR showed the highest standard deviation (4.72) and widest range (15 points), suggesting strong differentiation across papers. On the other hand, Grok-Fast exhibited lower variability (SD = 2.57, Range = 8) with less discrimination between articles.

Notably, Grok-Fast and Grok-Expert were based on the same underlying architecture but produced different scoring patterns. Grok-Expert showed lower variability (SD = 2.21, range = 7) and the highest average scores, while Grok-Fast produced scores closer to the overall sample mean but still showed similarly compressed scores (range = 8). This comparison suggests that Grok, regardless of approach, tended to produce compressed scores with limited differentiation across papers.

Analysis of dimension-specific patterns across AI systems revealed additional differences. Table 4 presents mean scores for each dimension broken down by AI evaluator, identifying which systems scored each dimension most generously or strictly.

Table 4. Mean Dimension Scores by Artificial Intelligence System across All Ten Papers

| AI System | Actionability | Novelty | Feas. Problem | Feas. Resources | Impact | Accessibility |
|---|---|---|---|---|---|---|
| Grok-Expert | 6.30 | 5.80 | 6.20 | 6.50 | 6.30 | 6.90 |
| ChatGPT4o DR | 5.00 | 5.90 | 4.60 | 5.10 | 5.20 | 6.80 |
| Opus 4.5 | 5.80 | 5.30 | 4.70 | 4.90 | 5.70 | 5.20 |
| Grok-Fast | 5.70 | 5.60 | 4.60 | 4.80 | 5.60 | 5.50 |
| Opus 4.5 ET | 4.70 | 5.20 | 4.10 | 3.80 | 5.00 | 6.10 |
| ChatGPT4o | 3.00 | 5.70 | 2.90 | 3.00 | 4.20 | 4.20 |
| Sample Mean | 5.08 | 5.58 | 4.52 | 4.68 | 5.33 | 5.78 |

*Note*. Each cell represents mean score across ten papers for each dimension and AI system.

Accessibility scores showed the highest means across most AI systems, with ChatGPT4o DR and Grok-Expert both averaging above 6.8. However, variation existed, with ChatGPT4o averaging only 4.2 on Accessibility compared to 6.9 for Grok-Expert, suggesting different definitions of what "accessible" research meant to each system.For Grok-Expert, its rationale was, "I rated 7 (Strongly agree) for all because the documents were readily available in the context of this rating task, assuming an organizational or evaluative setting where such provision implies accessibility," (xAI, 2025). For ChatGPT4o, its rationale was, "while their writing was generally clear and structured, many relied on academic terminology and theoretical framing that required interpretation by managerial readers," (OpenAI, 2025). These two responses support the claim above that these systems had different definitions.

Novelty demonstrated the most consistent scoring across AI systems, with means ranging only from 5.2 (Opus 4.5 ET) to 5.9 (ChatGPT4o DR), This consistency suggested that AI evaluators agreed about what constitutes novel research compared to other dimensions.

Actionability exhibited the widest variation across AI systems, with means spanning from 3.0 (ChatGPT4o) to 6.3 (Grok-Expert). This difference indicated disagreement across AI systems about what qualifies as "actionable." ChatGPT4o appeared to apply strict standards while Grok-Expert accepted more general frameworks as actionable.

Both Feasibility dimensions showed patterns that raise concerns. ChatGPT4o assigned mean scores at or below 3.0 for both Feasibility criteria, suggesting this system perceived most research as failing to solve genuine problems or requiring more steps or resources. In contrast, Grok-Expert averaged above 6.2 on both Feasibility dimensions, viewing the same papers as addressing real managerial challenges with reasonable resource requirements.

Impact scores demonstrated moderate variation across systems, ranging from 4.2 (ChatGPT4o) to 6.3 (Grok-Expert). The remaining four systems placed between 5.0 and

5.7, suggesting greater agreement about potential organizational benefits. This pattern showed that AI systems found it easier to agree about whether the research generated value.

Papers receiving consistently high scores (mean above 33 across all six AI systems) included Taheri Ruh et al. (2025), Bjørgum et al. (2021–2022), and Essuman et al. (2023). These papers achieved high scores from both generous systems like Grok-Expert and strict systems like ChatGPT4o, suggesting that regardless of the system, these papers could be viewed as having higher managerial relevancy.

### 2. Research Question 2

The second research question examined what challenges exist when using refined GPT models to score managerial implications in academic research. The multi-rater evaluation design used in this study, utilizing six different AI systems, provided evidence about both technical and conceptual challenges in AI relevance assessment. Analysis of scoring patterns, inter-rater variability statistics, and dimension disagreements revealed three main challenges that limit current AI evaluation capabilities.

#### *a. Challenge 1: Inter-Rater Variability Across AI Systems*

The most important challenge identified was the variability in how different AI systems scored identical papers on identical criteria using identical scoring scales. The 15-point difference in mean total scores between the most generous evaluator (Grok-Expert: M = 38.0) and most strict evaluator (ChatGPT4o: M = 23.0) is most shocking. This suggests that the AI model selected has a strong influence on the scores.

Individual papers demonstrated even more volatile inter-rater agreement. According to Table 1, Jain & Huang (2020) received total scores ranging from 17 to 37 across the six AI systems, a 20-point spread. Paruchuri et al. (2024) spanned from 20 to 36, Molitoris et al. (2025) spanned 27 to 40, and Pal & Altay (2019) spanned from 26 to 40.

The pattern was not limited to total scores. According to Table 4, Dimension-specific analysis revealed disagreement as well. On Actionability, the six AI systems' means ranged from 3.0 to 6.3. Feasibility dimensions spanned from 2.9 to 6.5 points. This

disagreement across dimensions indicated differences in how AI systems interpret evaluation criteria.

### *b. Challenge 2: Architecture Effects*

ChatGPT4o and ChatGPT4o DR both utilized GPT-4 architecture but differed in whether deep reasoning processes were engaged. According to Table 3, ChatGPT4o DR produced mean scores 9.6 points higher than ChatGPT4o (32.6 vs. 23.0).

Similarly, Grok-Fast and Grok-Expert shared the Grok architecture but differed. According to Table 3, Grok-Expert averaged 7.2 points higher than Grok-Fast (38.0 vs. 30.8). Opus 4.5 and Opus 4.5 ET both employed Claude Opus architecture, with extended thinking reducing mean scores by 2.7 points (31.6 vs. 28.9).

These significant score shifts indicated that AI evaluation depends heavily on which instantiation of the model a researcher selects. The same model produced dramatically different assessments depending on whether systems were instructed to engage deep reasoning or use extended processing time.

### *c. Challenge 3: No Convergence Even with Identical Instructions*

Each of the six AI systems got the exact same set of instructions for evaluating all six dimensions and the same 1 to 7 scoring rubric. Despite this standardization, the systems failed to converge on assessments. This showed that current AI systems have distinct underlying processes that become pronounced when they are asked to apply well-defined evaluation criteria to complex, subjective judgments about research quality.

## B. CONCLUSION OF FINDINGS

The Appendix provides a table that lists each article along with scores from all AI evaluation models, as well as the coefficient of variation and the related agreement category. Agreement levels were categorized based on CV thresholds: high agreement (CV $\leq$ 10%) reflects strong agreement across AI systems with minimal scoring variation, moderate agreement (CV 10–20%) reflects general alignment with some variation, and low agreement (CV > 20%) reflects substantial disagreement among AI systems (Reed et al.,

2002). The findings presented in this chapter provide evidence addressing both research questions through multi-rater AI evaluation of managerial implications in management and supply chain research. The evidence reveals challenges for both research quality assessment and AI evaluation reliability.

THIS PAGE INTENTIONALLY LEFT BLANK

# V. CONCLUSION

This exploratory study investigated the ability of artificial intelligence systems to evaluate managerial relevance in academic management research, addressing a problem within academia regarding the practical applicability of published work. This final chapter synthesizes the findings from the multi-AI evaluation experiment, offers recommendations for researchers, practitioners, and journal editors, acknowledges the limitations of the research design, and makes suggestions for future inquiry.

## A. SUMMARY OF FINDINGS

This study addressed two research questions examining artificial intelligence evaluation capabilities and managerial relevance in management scholarship. The first research question asked how well recent research papers published in top management and supply chain journals deliver strong managerial implications based on established criteria of Actionability, Novelty, Impact, Accessibility, Feasibility (Problem-Solving), and Feasibility (Not Resource Constrained). The second research question explored whether AI systems can produce reliable and consistent results when performing subjective evaluations that require careful judgment about research quality and practical relevance.

Regarding the first research question, the evaluation of ten papers published between 2015 and 2025 in the Journal of Management, Journal of Business Logistics, and Journal of Purchasing and Supply Management revealed that management and supply chain scholarship delivers moderately strong to strong managerial implications according to AI assessment. The aggregate sample mean across all papers and all six AI raters was 31.4 out of 42 possible points (SD = 3.78). This finding suggests that recent work in top management and supply chain journals maintains meaningful connections to practitioner concerns.

For the second research question on AI evaluation reliability, the study found several challenges that limit confidence in using current AI systems as dependable research assessors. Despite receiving identical evaluation instructions specifying the same six dimensions and the same seven-point scoring scale, the six AI systems failed to converge

on similar evaluations. The amount of variation showed that current AI systems cannot reliably apply clear evaluation criteria when they are asked to make complex, subjective judgments about research quality.

Overall, management and supply chain scholarship does appear to deliver meaningful managerial implications according to AI assessment. However, the variability across AI systems limits confidence. The current state of AI technology appears inadequate for reliable, standardized assessment of complex research attributes, though the technology may still offer value with carefully designed hybrid methods.

## B. RECOMMENDATIONS

The findings from this multi-AI evaluation study offer several recommendations for researchers, journal editors, and practitioners seeking to bridge the gap between academic management scholarship and managerial practice.

### 1. Practical Recommendation for Editors: Implement Hybrid Human–Artificial Intelligence Screening Systems with Explicit Procedures

Journal editors who are dealing with increased volume of submissions may see AI-assisted screening as appealing, but these findings suggest that AI should not be used as the only decision-maker in publication reviews. Instead, editors should use hybrid approaches where AI provides one source of input and human experts provide another, with humans remaining the final decision-making authority.

### 2. Practical Recommendation for Researchers: Explicitly Address Multiple Relevance Dimensions

Researchers who want their work to have more practical impact can use these findings as guidance. Since Feasibility received the lowest mean scores, it is an important area to strengthen. Authors should explain both how their recommendations address real problems and how managers can put them into practice without resource strain. Researchers who want their work to have practical impact should use plain language abstracts, provide executive summaries of the main findings, and include additional materials written for practitioners. The results of this study suggest that open access efforts

have reduced barriers, therefore, improving how research is presented is now the main area where accessibility can still be strengthened.

## C. LIMITATIONS

Although this study provides new insights into AI evaluation and managerial relevance in management research, it also includes several limitations that readers should keep in mind when interpreting the findings and planning future research.

### 1. Limitation I: Sampling Bias

The paper selection process introduced possible sampling biases that may limit how well the findings represent management scholarship as a whole. The papers came from journals in management, supply chain, logistics, and business research, chosen because they are influential in their fields. This intentional sample provided variation in managerial implications but left out large portions of management scholarship. Additionally, this study focused on papers published between 2015 and 2025. Papers from these years may differ in how they show relevance compared with earlier work. Additionally, including the *Journal of Management* alongside supply chain journals may have introduced bias. Prior research suggests that management studies tend to place greater emphasis on theory and rigor than on manager accessibility (Kieser & Leiner, 2009). Lower scores may reflect norms in management studies rather than a lack of practical value.

### 2. Limitation II: Small Sample Size

The evaluation of ten papers allowed for an exploratory study across AI systems, but it also limited statistical power. With only ten observations per AI system, the standard errors for mean scores and dimensional statistics remain relatively large, and the study cannot detect smaller effects. The data cannot reliably show whether AI system performance varies by journal or whether publication year shapes relevance assessments. These kinds of relationships may be meaningful, but they require much larger samples to analyze with confidence. Future research should use much larger samples.

### 3. Limitation III: Version Control Challenges

The AI systems used in this study reflect the specific versions that were available at the time of data collection, but these systems are updated frequently, and those updates can change how they evaluate papers. Because of this, researchers who try to replicate the study using the same systems may still get different results if updates have occurred. Future research should use strict version-tracking procedures and consider saving system outputs so that researchers can later examine how responses change across versions. This limitation shows that AI evaluation is constantly evolving. It should be treated as a dynamic method rather than a fixed tool.

### 4. Limitation IV: Single Prompt Without Iterative Refinement

This study used one standardized prompt that was given only once to each AI system, without any refinement or repeated prompting that might have improved evaluation quality or consistency. As a result, the single-prompt design may reflect lower AI performance than what could be achieved through iterative refinement. In addition, different AI systems may respond differently to repeated or refined prompts, with some models showing noticeable improvement and others showing little change (Reynolds & McDonell, 2021). This means that some of the variability observed across systems may be due to differences in how sensitive the models are to prompt iteration, rather than differences in their evaluation capabilities.

## D. AREAS OF FUTURE RESEARCH

This study created additional opportunities for future work that could deepen understanding of AI evaluation capabilities, the nature of managerial relevance in academic research, and methods for narrowing the research–practice gap. The following research suggestions stem from the findings and limitations documented in this paper.

### 1. Future Research I: Conduct a Similar Study with an Expanded Sample Size

The next step is to replicate the multi-AI evaluation design using a much larger sample. This replication would address the small sample size in the current study while

testing whether the observed patterns remain consistent across a broader range of management scholarship. Additionally, the study should use samples across journal tier levels to examine whether AI systems rate differently depending on journal prestige.

### 2. Future Research II: Investigate Defense Management Journals

This study examined only commercial management, supply chain, and business logistics research, but a parallel body of scholarship exists in defense-focused publications. These publications deserve similar analysis. Future studies should apply the multi-AI evaluation approach to these outlets to explore managerial relevance in defense scholarship and to determine whether patterns observed in commercial research remain true in this domain. A study of AI evaluations across defense management scholarship would make it possible to test whether the patterns identified in this paper span across sectors or come from domain-specific dynamics. Extending the analysis to defense management journals would also clarify whether AI systems accurately interpret defense-specific terminology, structures, and policies.

### 3. Future Research III: Examine the Effects of Iterative Refinement

Future research should examine how iterative refinement affects AI evaluation results and identify prompting strategies that improve reliability. Studies could also compare single-prompt approaches with iterative methods in which AI systems receive feedback on their initial evaluations and are allowed to revise their scores. This aligns with recent research on self-refinement in large language models (Madaan et al., 2023). Results from this work could help develop practical guidance on how to best use AI systems for research evaluation tasks.

## E. IN SUMMATION

In summation, this study has highlighted both the promise and the current limits of artificial intelligence as a tool for evaluating managerial relevance in academic management research.

Moving forward, progress will require additional methodological development. This includes using multiple AI systems, validating outputs against expert human

judgment, exploring fine-tuned, temperature adjusted models, and extending this research into fields such as defense management.

This paper provides one step in that progress by documenting the current state of AI evaluation capabilities and outlining suggestions needed to improve the practical impact of management scholarship.

# APPENDIX. ARTIFICIAL INTELLIGENCE EVALUATOR RATINGS, METRICS, AND AGREEMENT LEVELS BY PAPER

| Paper | Journal | ChatGPT4o | ChatGPT4o DR | Grok-Fast | Grok-Expert | Opus 4.5 | Opus 4.5 ET | Mean | SD | CV (%) | Agreement |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Developing and deploying competences for innovative public procurement *Taheriruh et al. (2025)* | JPSM | 28 | 38 | 35 | 41 | 37 | 34 | 35.5 | 4.55 | 12.8 | **Moderate** |
| Low power, high ambitions: New ventures *Bjørgum et al. (2021)* | JPSM | 31 | 37 | 35 | 40 | 36 | 31 | 35.0 | 3.58 | 10.2 | **Moderate** |
| In search of operational resilience *Essuman et al. (2023)* | JBL | 32 | 36 | 33 | 39 | 35 | 33 | 34.7 | 3.20 | 9.2 | **High** |
| Implementing PBC as procurement strategy *Molitoris et al. (2025)* | JPSM | 27 | 34 | 34 | 40 | 31 | 27 | 32.2 | 5.38 | 16.7 | **Moderate** |
| Identifying Key Success Factors BoP *Pal & Altay (2019)* | JBL | 26 | 34 | 34 | 40 | 32 | 26 | 32.0 | 5.10 | 15.9 | **Moderate** |
| Supply Chain Characteristics influence rival responses *Guntuka et al. (2024)* | JBL | 28 | 32 | 33 | 37 | 31 | 30 | 31.8 | 3.06 | 9.6 | **High** |
| Pack-and-a-Half Rule eliminate backroom inventories *Eroglu et al. (2018)* | JBL | 29 | 31 | 30 | 38 | 30 | 28 | 31.0 | 3.63 | 11.7 | **Moderate** |
| Governance Failure and Governance Under Failure | JOM | 20 | 33 | 28 | 36 | 29 | 20 | 27.7 | 6.44 | 23.2 | **Low** |

| Paper | Journal | ChatGPT4o | ChatGPT4o DR | Grok-Fast | Grok-Expert | Opus 4.5 | Opus 4.5 ET | Mean | SD | CV (%) | Agreement |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Paruchuri et al. (2024)* | | | | | | | | | | | |
| Oh the Anxiety! Supervisor Bottom-Line Mentality *Mitchell et al. (2023)* | JOM | 22 | 30 | 28 | 35 | 28 | 22 | 27.5 | 3.21 | 11.7 | **Moderate** |
| Learning from the Past: Scientist Relocation *Jain & Huang (2020)* | JOM | 17 | 30 | 28 | 37 | 25 | 21 | 26.3 | 6.83 | 26.0 | **Low** |

Note. Papers ordered by descending mean total score.

# LIST OF REFERENCES


Academy of Management. (n.d.). *Academy of Management Perspectives*. https://www.aom.org/publications/journals/perspectives/

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 715–725. https://aclanthology.org/P16-1068/

Ankers, P., & Brennan, R. (2002). Managerial relevance in academic research: An exploratory study. *Marketing Intelligence & Planning, 20*(1), 15–21. https://doi.org/10.1108/02634500210414729

Anthropic. (2025, February 24). Claude's extended thinking. https://www.anthropic.com/news/visible-extended-thinking

Arteaga, E., Biesbroek, R., Nalau, J., & Howes, M. (2024). Across the great divide: A systematic literature review to address the gap between theory and practice. *SAGE Open, 14*(1). https://doi.org/10.1177/21582440241228019

Bansal, P., Bertels, S., Ewart, T., MacConnachie, P., & O'Brien, J. (2012). Bridging the research–practice gap. Academy of Management Perspectives, 26(1), 73–92. https://doi.org/10.5465/amp.2011.0140

Bartunek, J. M., & Rynes, S. L. (2014). Academics and practitioners are alike and unlike: The paradoxes of academic–practitioner relationships. Journal of Management, 40(5), 1181–1201.

Bjørgum, Ø., Aaboen, L., & Fredriksson, A. (2021). Low power, high ambitions: New ventures developing their first supply chains. Journal of Purchasing and Supply Management, 27(3), Article 100670. https://doi.org/10.1016/j.pursup.2020.100670

Bland, J. M., & Altman, D. G. (1996). Statistics notes: Measurement error. BMJ, 313(7059), 744. https://doi.org/10.1136/bmj.313.7059.744

Cao, Z., Li, M., & Pavlou, P. A. (in press). AI in business research. *Decision Sciences*. https://doi.org/10.2139/ssrn.4897619

Carton, G. and Mouricou, P. (2017). Is management research relevant? A systematic analysis of the rigor–relevance debate in top-tier journals (1994–2013) M@n@gement, *20*(2), 166–203. https://doi.org/10.3917/mana.202.0166

Chai, C. S., Wang, X., & Xu, C. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Heliyon, 10*(3), e25352.

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications, 8*(1), 1–11. https://doi.org/10.1057/s41599-020-00703-8

Dasgupta, T., Naskar, A., Dey, L., & Saha, R. (2018). Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, 93–102. https://aclanthology.org/W18-3713/

Daugherty, P. R., & Wilson, H. J. (2024). Embracing Gen AI at work. *Harvard Business Review*, *102*(5), 151–155. https://www.accenture.com/content/dam/accenture/final/capabilities/technology/technology-innovation/document/Accenture-HBR-Embracing-Gen-AI-at-Work-Sept-Oct-2024-COVER-STORY.pdf

Emerald Publishing. (n.d.). Scope. *Supply Chain Management: An International Journal*. Retrieved November 6, 2025, from https://www.scimagojr.com/journalsearch.php?clean=0&q=23644&tip=sid

Eroglu, C., Williams, B. D., & Waller, M. A. (2018). Using the pack-and-a-half rule to eliminate backroom inventories in retail operations. Journal of Business Logistics, 39(3), 164–181. https://doi.org/10.1111/jbl.12191

Essuman, D., Ataburo, H., Boso, N., Anin, E. K., & Appiah, L. O. (2024). In search of operational resilience: How and when improvisation matters. Journal of Business Logistics, 45(2), Article e12312.

Evans, M. (2025, November 6). *Choosing AI, human research—or both?* Insight into Academia. https://insightintoacademia.com/choosing-ai-human-research-or-both/

Fuller, M., & Bixby, J. (2024). The theoretical and practical implications of OpenAI system rubric assessment and feedback on higher education written assignments. *American Journal of Educational Research, 12*(4), 208–216. https://pubs.sciepub.com/education/12/4/4/index.html

Ge, L., Agrawal, R., Singer, M., Eltzschig, H. K., & Tian, J. (2024). Leveraging artificial intelligence to enhance systematic reviews in health research: Advanced tools and challenges. *Systematic Reviews*, *13*, Article 269. https://doi.org/10.1186/s13643-024-02682-2

Guntuka, L., Cantor, D. E., Corsi, T. M., D'oria, L., & Grover, A. (2025). Do supply chain characteristics influence a rival firm's responses to a focal firm's product preannouncements? A competitive dynamics perspective. Journal of Business Logistics, 46(1), Article e12395. https://doi.org/10.1111/jbl.12395

Hawkins, T. G., Gravier, M. J., Farris, M. T., Niranjan, S. & Ekezie, U. (2022). Exploring the impact of logistics and supply chain management scholarship: Why pursue practical relevance and are we successful? *Journal of Business Logistics*, *43*, 654–678. https://doi.org/10.1111/jbl.12306

Hawkins, T. G., & Muir, W. A. (2024). Leveraging a rubric for evaluating managerial relevance in logistics research. *Journal of Business Logistics, 45*(2), 134–152.

Hedgepeth, S., & Tagatac, R. M. (2023). Assessing DoD confidence and bias in AI/LLM authored evaluation factors (NPS-AM-24-015). Naval Postgraduate School, Acquisition Research Program. https://dair.nps.edu/handle/123456789/5034

Hoffman, A. J. (2015, November 23). Isolated scholars: Making bolder choices to make scholarship matter. *The Chronicle of Higher Education*.

Horbach, S. P. J. M., & Halffman, W. (2020). Innovating editorial practices: Academic publishers at work. *Research Integrity and Peer Review, 5*(1), 12. https://doi.org/10.1186/s41073-020-00097-w

Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review, 8*(1), 4. https://doi.org/10.1186/s41073-023-00133-5

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5*, e208. https://doi.org/10.7717/peerj-cs.208

Jack, A., & Dalal, A. (2025, November 1). Wharton tops FT ranking of business schools with impact. *Financial Times*. https://www.ft.com/content/8d9c9830-f2c3-4229-8762-87912d3a5e81

Jain, A., & Huang, K. G. (2022). Learning from the past: How prior experience impacts the value of innovation after scientist relocation. Journal of Management, 48(3), 571–604. https://doi.org/10.1177/0149206320979658

Jaworski, B. J. (2011). On managerial relevance. *Journal of Marketing, 75*(4), 211–224. https://doi.org/10.1509/jmkg.75.4.211

Kaldaras, L., Haudek, K. C., Wang, J., & Long, T. (2022). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Frontiers in Education, 7*, 983055. https://doi.org/10.3389/feduc.2022.983055

Kerr, S. (1975). On the folly of rewarding for A while hoping for B. *Academy of Management Journal, 18*(4), 769–783.

Kieser, A., & Leiner, L. (2009). Why the rigour–relevance gap in management research is unbridgeable. *Journal of Management Studies, 46*(3), 516–533. https://doi.org/10.1111/j.1467-6486.2009.00831.x

Kieser, A., Nicolai, A., & Seidl, D. (2015). The practical relevance of management research: Turning the debate on relevance into a rigorous scientific research program. *Academy of Management Annals, 9*(1), 143–233. https://doi.org/10.5465/19416520.2015.1011853

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Koraishi, O. (2024). The Intersection of AI and Language Assessment: A Study on the Reliability of ChatGPT in Grading IELTS Writing Task 2. Language Teaching Research Quarterly, 43, 22–42.

Kristof, N. (2014, February 15). Professors, we need you! *The New York Times*. https://www.nytimes.com/2014/02/16/opinion/sunday/kristof-professors-we-need-you.html

Kruse, B. G., Didion, J. L., & Perzynski, K. (2020). Improving student feedback quality: A simple model using peer review and feedback rubrics. *Medical Science Educator, 30*(3), 1105–1111.

Levman, J., Ewenson, B., Apaloo, J., Berger, D., & Tyrrell, P. N. (2023). Error consistency for machine learning evaluation and validation with application to biomedical diagnostics. *Diagnostics*, *13*(7), 1315. https://doi.org/10.3390/diagnostics13071315

Liang, W., Zou, A., Wu, T., Fu, D., Zhang, H., Naik, R., Wu, E., Narayan, A., & Zhang, P. (2023). *Can large language models provide useful feedback on research papers? A large-scale empirical analysis*. ArXiv. https://arxiv.org/abs/2310.01783

Liew, P. Y., & Tan, I. K. T. (2024). On automated essay grading using large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence (CSAI '24)* (pp. 204–211). Association for Computing Machinery. https://doi.org/10.1145/3709026.3709030

Lin, J. (2025). Large language models for automated scholarly paper review: A survey. *Information Fusion*, *124*, Article 103332. https://doi.org/10.1016/j.inffus.2025.103332

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), Advances in Neural Information Processing Systems 36 (pp. 46534-46594). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html

Mitchell, M. S., Hetrick, A. L., Mawritz, M. B., Edwards, B. D., & Greenbaum, R. L. (2024). Oh the anxiety! The anxiety of supervisor bottom-line mentality and mitigating effects of ethical leadership. Journal of Management, 50(7), 2885–2926. https://doi.org/10.1177/01492063231196553

Molitoris, C., Glas, A. H., & Eßig, M. (2025). Implementing PBC as a procurement strategy in the public sector: Understanding strategy-structure fit. Journal of Purchasing and Supply Management, 31(1), Article 100955.

Mostafapour, M., Fortier, J. H., Pacheco, K., Murray, H., & Garber, G. (2024). Evaluating literature reviews conducted by humans versus ChatGPT: Comparative study. *JMIR AI*, *3*, e56537. https://doi.org/10.2196/56537

Nicolai, A., & Seidl, D. (2010). That's relevant! Different forms of practical relevance in management science. *Organization Studies, 31*(9–10), 1257–1285. https://doi.org/10.1177/0170840610374401

Nobel, C. (2016, October 3). Why isn't business research more relevant to business practitioners? *Harvard Business School Working Knowledge*. https://www.library.hbs.edu/working-knowledge/why-isn-t-business-research-more-relevant-to-business-practitioners

OpenAI. (2025, February 2). Introducing deep research. https://openai.com/index/introducing-deep-research/

OpenAI. (2025, December 15). Accessibility score rationale for Jain & Huang (2020) [AI-generated response]. ChatGPT.

OpenAI. (2025). ChatGPT conversation on research relevance and accessibility scoring (ChatGPT4o). https://chat.openai.com/

Pal, R., & Altay, N. (2019). Identifying key success factors for social enterprises serving base-of-pyramid markets through analysis of value chain complexities. Journal of Business Logistics, 40(2), 161–179. https://doi.org/10.1111/jbl.12212

Pannucci, C. J., & Wilkins, E. G. (2010). Identifying and avoiding bias in research. *Plastic and Reconstructive Surgery, 126*(2), 619–625. https://doi.org/10.1097/PRS.0b013e3181de24bc

Paruchuri, S., Hoempler, E. A., Cowen, A. P., Cannella, A. A., Jr., & Nahm, P. I. (2024). Governance failure and governance under failure: Reviewing the role of directors in organizational misconduct. Journal of Management, 50(6), 2237–2265. https://doi.org/10.1177/01492063231225420

Phys.org. (2023, October 4). *Large language models show potential to assist in peer review, but human expertise remains essential*. https://phys.org/news/2023-10-large-language-peer-review.html

Poličar, P. G., Špendl, M., Curk, T., & Zupan, B. (2025, January 24). *Automated assignment grading with large language models: Insights from a bioinformatics course*. ArXiv. https://doi.org/10.48550/arXiv.2501.14499

Radović, S. (2024). Introduction to the SRL-S rubric for evaluation of innovative higher educational technology for self-regulated learning. *Innovative Higher Education, 49*(5), 903–935. https://doi.org/10.1007/s10755-024-09771-z

Reed, G. F., Lynn, F., & Meade, B. D. (2002). Use of coefficient of variation in assessing variability of quantitative assays. Clinical and Diagnostic Laboratory Immunology, 9(6), 1235–1239. https://doi.org/10.1128/cdli.9.6.1235-1239.2002

Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (Article 314, pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3411763.3451760

Sarker, I. H. (2022). AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, *3*(2), Article 158. https://doi.org/10.1007/s42979-022-01043-x

Schauerte, N., Becker, M., Imschloss, M., Wichmann, J. R., & Reinartz, W. J. (2023). The managerial relevance of marketing science: Properties and genesis. *International Journal of Research in Marketing, 40*(4), 801–822.

Setkute, J., & Dibb, S. (2025). From theory to practice: Practical implications as a translational bridge between research relevance and impact. *Industrial Marketing Management*, *125*, 131–149. https://doi.org/10.1016/j.indmarman.2024.12.017

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1882–1891. https://aclanthology.org/D16-1193/

Taheriruh, M., Jääskeläinen, A., Loijas, K., & Harrison, D. (2025). Developing and deploying competences for innovative public procurement: A network perspective. Journal of Purchasing and Supply Management, 31(3), Article 101039. https://doi.org/10.1016/j.pursup.2025.101039

Thelwall, M. (2023). Predicting article quality scores with machine learning: The U.K. research excellence framework. *Quantitative Science Studies, 4*(2), 547–563.

Thelwall, M. (2024). Can ChatGPT evaluate research quality? *Journal of Data and Information Science, 9*(2), 1–20. https://doi.org/10.2478/jdis-2024-0013

Thompson, S. K. (2012). *Sampling* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118162934

Van de Ven, A. H. (2007). *Engaged scholarship: A guide for organizational and social research*. Oxford University Press.

Varadarajan, R. (2020). Relevance, rigor and impact of scholarly research in marketing: State of the discipline and outlook. *AMS Review, 10*, 199–205. https://doi.org/10.1007/s13162-020-00180-x

Wang, S., Zhao, Y., Ouyang, D., & Wen, J. (2024). Using GPT-4 to write a scientific review article: A pilot evaluation study. *BioData Mining, 17*(1), 9. https://doi.org/10.1186/s13040-024-00371-3

Williams, R., & Grant, J. (2023). Exploring the application of machine learning to expert evaluation of research impact. *Research Evaluation, 32*(3), 389–401.

xAI. (2025). Grok (Version 4) response to "Why do you believe these papers are accessible?" [Large language model]. https://x.ai/grok

xAI. (n.d.). Reasoning. https://docs.x.ai/docs/guides/reasoning

Zhang, L., Huang, Y., & Chen, S. (2024). Evaluating large language models for criterion-based grading from agreement to consistency. *NPJ Science of Learning, 9*(1), 42. https://doi.org/10.1038/s41539-024-00291-1


Acquisition Research Program
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net