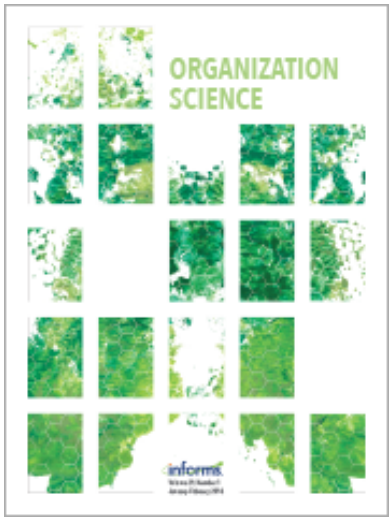


Organization Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review

Claudine Gartenberg, Sharique Hasan, Alex Murray, Lamar Pierce

To cite this article:
Claudine Gartenberg, Sharique Hasan, Alex Murray, Lamar Pierce (2026) More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review. Organization Science 37(3):795-812. https://doi.org/10.1287/orsc.2026.ed.v37.n3

https://pubsonline.informs.org/journal/orsc

ORGANIZATION SCIENCE
Vol. 37, No. 3, May–June 2026, pp. 795–812
ISSN 1047-7039 (print), ISSN 1526-5455 (online)

More Versus Better: Artificial Intelligence, Incentives, and the Emerging Crisis in Peer Review

Claudine Gartenberg,[a] Sharique Hasan,[b] Alex Murray,[c] Lamar Pierce[d,*]

[a] The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104; [b] Duke University, Durham, North Carolina 27708; [c] University of Oregon, Eugene, Oregon 97403; [d] Washington University in St. Louis, St. Louis, Missouri 63130

*Corresponding author

Contact: cgart@wharton.upenn.edu, https://orcid.org/0000-0003-0696-7266 (CG); sharique.hasan@duke.edu, https://orcid.org/0000-0002-8574-8610 (SH); amm16@uoregon.edu, https://orcid.org/0000-0001-6161-8363 (AM); pierce@wustl.edu, https://orcid.org/0000-0003-0452-1391 (LP)

Received: April 2, 2026
Revised: April 2, 2026
Accepted: April 2, 2026
Published Online in Articles in Advance: April 27, 2026

https://doi.org/10.1287/orsc.2026.ed.v37.n3

Abstract. As the AI Task Force for *Organization Science*, we provide an early account of artificial intelligence's (AI) impact on both submissions and reviews at a major academic journal. Submission volume has risen 42% since the late 2022 release of ChatGPT, while writing quality has declined. The rise in AI-generated writing accounts for nearly all of these trends. AI-generated writing in reviews has also increased, and is characterized by lower writing quality and less topical diversity than human-generated writing. We are, to our knowledge, the first journal to report these early impacts of AI in the review process. Conversations with editors across scientific disciplines, however, suggest that what we observe is not limited to our journal or to the social sciences. At this early stage of AI adoption, we cannot make a normative assessment about appropriate or ideal levels of AI usage. We can, however, conclude that the current state of AI tools, amplified by existing publish-or-perish incentives, appears to be pushing the system toward an equilibrium of more rather than better research. Reaching an equilibrium in which AI serves as a critical engine of innovation will require that our institutions and the incentive structures they create adapt.

Funding: S. Hasan used research funding from Duke University's Fuqua School of Business. C. Gartenberg used research funding from University of Pennsylvania's Wharton School.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/orsc.2026.ed.v37.n3.

Keywords: Artificial Intelligence • Peer Review • Science • Innovation • Evaluation • Incentives • Generative AI

1. Introduction

As the AI Task Force for *Organization Science*, we spent the last several months studying how generative artificial intelligence (AI) is affecting the workings of our journal. Our goal was not to cast judgment on what defines appropriate AI use, but rather to provide evidence on how growing AI usage might be reshaping how we evaluate and promote scientific research. A clear picture is emerging: AI language models, combined with strong publish-or-perish incentives, are pushing our field to produce more rather than better research. Journal operations are increasingly strained by growth in low-quality AI-augmented submissions, affecting volunteer editors and reviewers alike. While most of these submissions are rejected, many during the initial deputy editor screen, the burden they impose is substantial. We believe that a better equilibrium is possible, one in which we work with AI to produce research once thought impossible. Reaching this equilibrium will require more than just better tools, however; it will require better institutions.

In this report, we analyze how submission and review activity at *Organization Science* have changed following the release of ChatGPT, the first of many modern commercial large language models (LLMs), in November 2022. We do so using the full corpus of 6,957 initial submissions and 10,389 text-inputted reviews handled by the journal since 2021.[1] Focusing on the peer review pipeline provides a distinctive perspective on AI's effects on scientific research. Studies of published articles conflate authors' contributions with editorial selection and revision processes, whereas we can separate these. And, because submitting to *Organization Science* incurs real costs in terms of elapsed time under review, these submissions reflect authors' beliefs about quality. Most importantly, our access to reviewer comments and editorial outcomes allows us to observe how human judgment—the scarce input into the review process—responds to the rise of AI in submissions and reviews.

We report AI usage over a five-year window beginning in January 2021, with the first two years serving as our "placebo" period for AI writing. We calculate an AI writing score for each submission and review using Pangram, a widely used and validated AI detection tool. On the submission side, we use abstracts as our primary text, and validate our patterns using a smaller sample of full manuscripts to ensure that AI use in abstracts is, on

average, representative of its use in full submissions. Finally, we use the 79% of reviews submitted in text format (rather than as PDFs).[2]

Although Pangram is regarded as the most accurate AI detection tool available (Jabarian and Imas 2025), no statistical tool is fully reliable when applied to individual texts. Our analysis is structured accordingly: we report only aggregate trends across thousands of submissions and reviews, avoiding focus on individual submissions and reviews. Also, rather than applying a binary threshold to classify texts as AI-generated or human-written, we use Pangram's continuous score to track shifts in estimates of AI prevalence over time. Finally, we include a two-year placebo period prior to ChatGPT's release to establish a baseline against which to measure change. Together, these choices allow us to identify meaningful shifts in the prevalance and quality of AI writing, even as individual-level detection remains imperfect.

What we find is sobering. Submission volume has risen by 42% since November 2022. At the same time, submission writing quality also began to decline at the end of 2022, with Flesch Reading Ease (a standard measure of writing quality) 1.28 standard deviations (SD) lower in January 2026 relative to January 2021. Submissions that are heavily AI-generated account for nearly all of these trends. The quality of reviews has also dropped sharply since November 2022, driven by an increase in heavily AI-generated reviews. These reviews, in addition to being of worse quality, are also narrower in their emphasis, focusing more on theory and less on data. In short, *more* research is being submitted, with *more* AI writing that is lower in quality—*not better*.

These trends put pressure on a system already under stress. Even before ChatGPT's introduction, submission volumes were climbing, driven by high-powered incentives such as publication-count criteria for tenure and financial bonuses. Journals that appear on lists like the UT-Dallas (UTD) and Financial Times 50 (FT50), face particular pressure, as these lists define the "right" journals for many schools. If these count incentives are indeed fueling the post-ChatGPT rise in AI-augmented submissions, the effect should be stronger at institutions where these incentives are most pronounced. This is what we find. We proxy for incentive intensity using institutional responses to the launch of the UTD journal list in 2005. Schools whose publishing patterns shifted following the launch—which we designate as "UTD Responder" schools—exhibit a greater increase in AI-writing in submissions after ChatGPT's release.

We believe that AI has the potential to transform research and our field. The findings of our task force, however, suggest that we are not yet fulfilling this potential:

- First, the recent rise in journal submission quantity is primarily due to AI use, not organic growth in the field or increased journal reputation.
- Second, while many believe that AI improves writing quality, we show that AI has led to a deterioration of prose in submitted manuscripts and reviews.
- Third, we show that the authors who are most likely to use AI in writing—research teams from non-native English-speaking institutions and new entrants to the field—do not benefit in the review process and, in fact, may be hurt when using AI for writing.
- Fourth, while prior research suggests a strong overlap between AI and human reviews, we show that AI reviews are much narrower than human ones.
- Finally, and perhaps most importantly, celebrating AI as a productivity-generating technology misses its collision with the field's incentive structures. When institutional incentives push "pubs" over rigorous research, AI will be used to drive volume rather than to fulfill its potential as a critical research technology.

These findings suggest a complex, systemic challenge with no easy answer for journals, authors, or universities. One point is clear to us: AI has already entered academia with gusto. Based on conversations with editors, both inside and outside our field, the patterns we reveal at *Organization Science* are surely pervasive across the sciences (see also Naddaf 2025, Jones 2026, Munger et al. 2026). Trends in "vibe-researching" are increasingly discussed on social media. NeurIPS, the premier AI conference, saw a 75% increase in research submissions in 2025 compared with 2023, swamping its human review system and leading to AI-generated research being accepted into the program. Kusumegi et al. (2025) examine preprint production and find that authors who adopt AI upload 36% more preprints than nonadopters across the three major preprint servers (arXiv, bioRxiv, and SSRN). Their study period ended in June 2024, and the trends have continued and now affect actual journal submissions and likely publications.

If our findings are, in fact, widespread beyond *Organization Science* and if they signify AI usage beyond what we can observe in writing detectors, they point to a systemic need to reconsider institutions beyond academic journals, including how we train students, incentivize faculty, and judge new ideas and the people developing them. We are not alone in this conclusion (Bechky and Davis 2025). Our response cannot be purely technical approaches or consist solely of disclosure policies or AI "bans." Disclosure does not solve the volume problem, and AI bans are both unrealistic and counterproductive to research progress. The challenge, therefore, is first to understand what is happening and then to steer the equilibrium toward better rather than just more research.

In the end, academia requires engaged, expert judgment from a community of scholars who invest in evaluating each other's work. If AI only amplifies volume—and lowers quality in the process—we may end up with more research but less collective

knowledge. As editors who serve as the bulwark against inundating reviewers with low-quality AI-driven submissions, we view the current system as unsustainable. As our near-universal rejection rates of AI-driven writing show, our human effort has so far protected the review process from this tide. But the humans are getting tired.

2. Data and Measures

We analyzed AI usage across all *Organization Science*'s initial (first-round) submissions and text-inputted reviews from January 2021 to February 2026, providing data on journal operations before and after the launch of ChatGPT in November 2022. For submissions, we focus primarily on abstracts, using the full texts of a subset of manuscripts to evaluate whether AI use in abstracts is a reasonable proxy for AI use in the full texts. For peer reviews, we use the full text of all referee reports submitted in text format directly into the ScholarOne system (as opposed to uploaded as PDF), which accounts for 79% of all reviews. In total, this consists of 6,957 submissions by 11,887 authors, reviewed 10,389 times by 2,519 unique reviewers.[3]

2.1. AI Detection

Each text is scored by Pangram (v3.1), a deep learning model trained to classify text as human- or AI-generated. The model is trained with asymmetric loss weighting that makes false positives (human text flagged as AI) five times less likely than false negatives (AI text classified as human). Pangram reports a 99.84% accuracy rate on human-written text,[4] repeatedly updating the models as new AI generations are released. Pangram cannot detect AI-usage in nonwriting tasks such as background research or coding, nor can we make any claims about this usage.

A 2025 independent AI detection tool evaluation by economists Jabarian and Imas (2025) found Pangram to have a false positive rate of less than 0.001 when Pangram evaluated samples written by GPT-4.1, Claude Opus 4, Claude Sonnet 4, and Gemini 2.0 Flash models, with false negative rates between 0.0045 and 0.038. We are using a more advanced Pangram model updated since the most recent AI releases (e.g., Opus 4.6), primarily applied to historical submissions created under older technologies, which strengthens our confidence in its accuracy. Even if the most recent papers we analyze used new advanced models, and the latest Pangram update does not keep up (for which we have no evidence), the model is trained to minimize false positives (human writing labeled as AI). Given that dynamic, we might, if anything, understate the prevalence of AI writing in 2026 submissions and reviews.

The model outputs a continuous "AI assistance score" ranging from zero to one, representing the estimated degree of AI pervasiveness in the text. Zero indicates fully human-written, and one indicates fully AI-generated. Pangram segments text into approximately 400-word windows and scores each independently. For multisegment texts (manuscripts), we compute a word-count-weighted average across segments to approximate the proportion of text exhibiting AI characteristics. For single-segment texts (abstracts), which typically fall within a single window, the score blends the model's confidence that AI was used with its estimate of how much of the text reflects AI involvement. The score should therefore be interpreted as an overall signal of AI prevalence in writing rather than the precise proportion of AI-generated text.

We use both this continuous score and also classify abstracts into four mutually exclusive categories. We define a score below 15% as having little to no AI writing use by the authors. A score between 15%–30% represents a more "collaborative" use of AI in writing, where the researcher retains primary creative effort. Scores between 30% and 70% cross the threshold from primary human ownership to giving the AI substantially more control. Finally, we view scores above 70% as primarily AI-generated content. These thresholds are judgment calls, but different categorizations produce similar findings.[5] This editorial, for example, received a Pangram score of 8.8, likely reflecting the light AI-based copyediting disclosed at the beginning. We do not use these scores to define an "optimal" or "appropriate" level of AI-usage in writing.

It is important to make clear that under the data processing agreement, Pangram deleted all data after evaluation and did not use them to train its general model, thereby complying with the INFORMS privacy policy.

2.2. Writing Quality

We calculate a set of standard writing-quality measures for each submission, including Flesch Reading Ease, Flesch-Kincaid Grade Level, FOG Index, SMOG Index, Nominalization, Passive Voice, Jargon, Hedging, and Specificity. Online Appendix A4 provides a detailed description of the scores calculations, which capture two broad categories of information about writing quality: (a) readability, the extent to which a text is easier or harder to read for the average person; and (b) style, the extent to which certain stylistic features appear in the text (e.g., nominalizations—where verbs and adjectives are converted into nouns, passive voice, or hedging with words such as "might", "could", and "suggest").

2.3. Review Topics and Emphasis

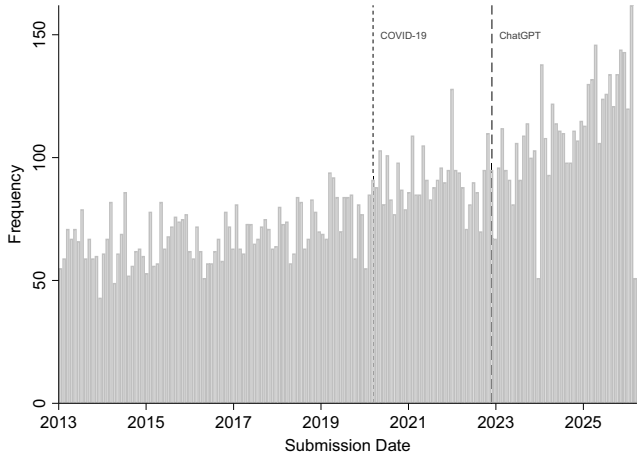
To understand what topics reviewers raise in their reports, we construct a simple measure of emphasis in reviews. For each report, we count the frequency of specific words indicating five broad foci: theory, contribution, clarity, data, and empirics. Theory words

include: theory, theoretical, theorize, mechanism, framework, construct, proposition, etc. Contribution words include research gap, implications, significance, and contribution. Clarity words include readable, well-written, rewrite, confusing, etc. Data words include data, sample, experiment, panel, longitudinal, etc. And finally, words related to empirics include regression, estimate, exogenous, endogenous, and identification. A full list of words is in Online Appendix A5. The word counts are then converted to relative shares to account for variation in the length of the review. To illustrate, the variable "theory" is the relative emphasis on theory in the review versus "data" or "empirics."[6]

2.4. Author and Manuscript Information and Outcomes

We supplement these AI measures with author team and reviewer attributes. Specifically, we focus on two author-team metrics. First, we construct a variable, *All non-Native English Institution Authors*, that equals 1 if all authors are based at institutions in nonnative English-speaking countries (e.g., not based in the United States, Canada, United Kingdom, Australia, South Africa, or New Zealand) and 0 otherwise. On average, we anticipate that such teams are less likely to include native English speakers and more likely to view AI as both a language and a writing assistant. This is clearly a noisy measure, as authors at such institutions may be native speakers or excellent writers, and vice versa. Similarly, some locations, such as Hong Kong, have multiple official languages, including English. Second, we construct a variable, *Prior OS Submissions*, that counts the number of submissions to *Organization Science* made by members of the author team over the past 15 years.

Figure 1. Monthly Submission Volume at *Organization Science* from January 2013 Until the End of 2025



Note. The dashed vertical line marks the launch of ChatGPT in November 2022 and the start of the COVID-19 pandemic in March 2020.

Finally, because we have full transaction data for each manuscript, we connect manuscripts to reviews and evaluate whether manuscript and review characteristics (e.g., AI use or author background) affect decisions and outcomes (e.g., desk rejects or rejections after review). A table of descriptive statistics is available in Online Appendix Table A1.

3. AI in the Submission Process

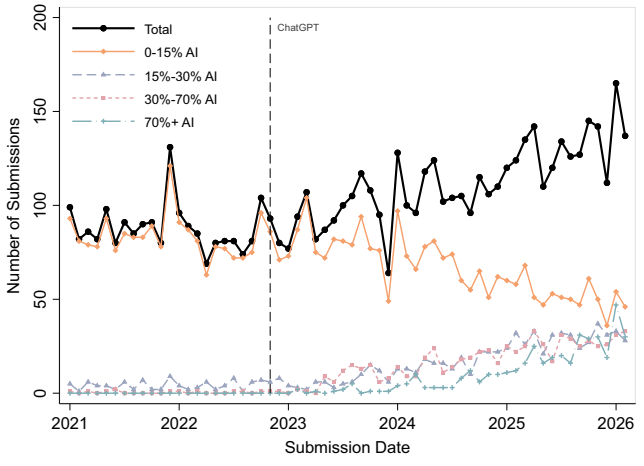
3.1. AI Use in Submissions is Increasing

We begin by analyzing submission volume over time at *Organization Science*. Like many journals, we have observed a substantial increase in submissions since late 2022. Figure 1 shows this rise. For instance, if we compare submission volume between January 2022–November 2023 (before ChatGPT, but after peak-COVID) to January 2024–November 2025 (after), we see 42% more submissions. In comparison, the COVID-19 bump in submissions was 20%, half the size of the AI effect.

While there could be many reasons for the rise in submissions, including reduced backlogs, increased scholar productivity, or journal reputation, Figure 2 suggests that the disproportionate increase in submission volume is driven by AI use. Post-ChatGPT, we see a marked decline in submissions flagged at 0%–15% AI (little to no AI use) and a corresponding rise in all other categories that make up the difference between the decline in human-only submissions and the 42% increase in total submissions.

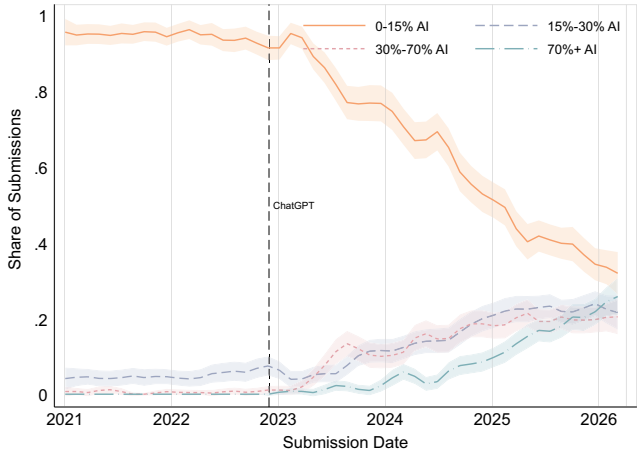
Figure 3 shows the shift in composition more clearly. It plots the relative shares of human- and AI-authored abstracts over time. Prior to the launch of ChatGPT, relative shares were flat. Nearly all submissions were classified as human (with some idiosyncratic noise).

Figure 2. Monthly Submission Volume by AI Use Categories over Time



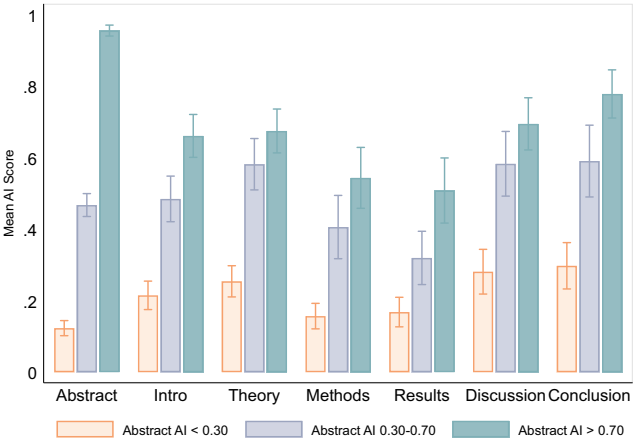
Notes. Each line shows the share of first submissions' abstracts classified as 0%–15% AI, 15%–30%, 30%–70%, 70%+. The dashed vertical line marks the launch of ChatGPT in November 2022.

Figure 3. Trends in AI Use Categories over Time



Notes. Each line shows the share of first-submission abstracts classified as 0%–15%, 15%–30%, 30%–70%, or 70%+ AI. The dashed vertical line marks the launch of ChatGPT in November 2022.

Figure 4. AI Scores for Each Section of a Sample of Manuscripts Stratified by AI Scores Falling into Low (<30), Medium (between 30 and 70), and High (>70) Categories



Note. Errors are larger for lower AI manuscripts because we oversampled on high AI for this analysis due to computational constraints.

Immediately after the launch of the first commercial LLM chatbots, a precipitous decline in human-only submissions began. At the same time, we observe a steady rise in all categories of AI-supported or generated submissions. What is most striking is that by February 2026, the majority of submissions submitted to *Organization Science* use AI in their writing to some degree. The most striking trend is the rise of the 70%+ AI category, where text is mostly or entirely generated by AI. While we are reporting AI use in abstracts, this trend is also evident in the subsample of full manuscripts we analyzed.

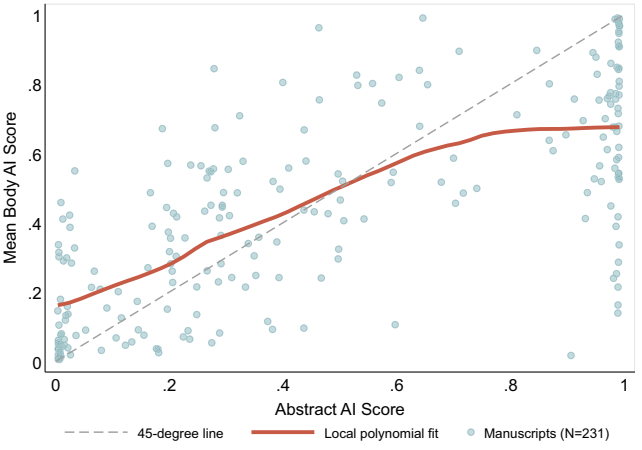
3.2. AI in Abstracts Reflects AI in Full Manuscripts

The prior graphs use only the abstract to compute a submission's AI score. We therefore tested whether AI use in the abstract is a reasonable proxy for AI use in the full manuscript. To do so, we examine the distribution of AI scores in a strategically sampled population of 230 manuscripts, namely those that score very high on AI and a subset of counterfactual manuscripts with medium and low AI scores. We focused on three sets of manuscripts: (a) 104 that were flagged as below 30%, (b) 49 that were flagged as between 30%–70%, and (c) 77 that were flagged at 70%+ AI. Due to computational and cost constraints, we limited our analysis to this sample. Group (a) had a mean abstract score of 0.123; (b) of 0.467; and (c) of 0.955.

We divide the manuscript into six sections: introduction, theory, methods, results, discussion, and conclusion, and then calculate Pangram scores for each section.

Overall, as shown in Figure 4, the full-text analysis of sampled manuscripts indicates that high AI scores (>70%) in the abstract do indeed signal high AI use throughout the manuscript. In particular, section-level Pangram scores show elevated AI usage across all sections, including introduction, theory, methods, results, discussion, and conclusion. AI scores are especially pronounced in writing-heavy sections such as the introduction, theory, discussion, and conclusion, but there is also meaningful AI use in the methods and results sections. In contrast, manuscripts classified as moderate (between 30% and 70%) and low AI (below

Figure 5. Scatterplot of AI Use in Abstract (x-axis) vs. AI Use Within the Manuscript Body (Mean of AI Use in Introduction, Theory, Methods, Results, Discussion, Conclusion)



Notes. High AI use in the abstract is highly correlated with AI use throughout the manuscript, but there is considerable variation at each level. Some manuscripts with low AI use in the abstract still have high AI use in the body, while some with high AI use in the abstract have much lower AI use in the paper.

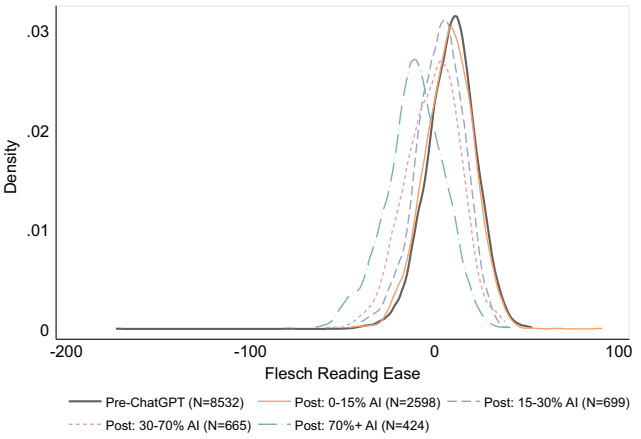
30%) based on the abstract show lower AI usage in the abstract but still indicate high levels of AI usage throughout, with similar patterns of relatively higher usage in more writing-heavy sections.

Overall, we feel confident that, in aggregate, AI use in abstracts is a reasonable proxy for AI use in manuscripts. However, we cannot make this assertion for any specific manuscript without directly measuring its full text. Indeed, Figure 5 suggests that while AI use in the abstract is a strong proxy for AI use throughout the manuscript, there is considerable variation across manuscripts.

3.3. AI is Associated with Worse Writing

So far, our analysis shows a steady rise in submissions, with AI-supported manuscripts dominating. We now turn to the central question: *Are these manuscripts higher-quality?* Figure 6 plots normalized Flesch Reading Ease for every initial submission to *Organization Science* from 2013 to February 2026, standardized to a mean of zero and standard deviation of 1. We do not find much evidence that the writing quality of those manuscripts changed meaningfully between 2013 and November 2022, when ChatGPT was launched. In contrast, post-ChatGPT, we see a precipitous decline in the average manuscript's Reading Ease score. Indeed, AI scores and Flesch Reading Ease are negatively correlated ($\rho = -0.4; p \leq .001$). While the prior figure reflects a significant decline in the mean reading ease across manuscripts, Figure 7 shows a marked shift in the distribution of reading ease scores across AI levels. Higher AI levels are associated with a negative shift in reading ease scores.

Figure 7. Distribution of Flesch Reading Ease by Level of AI in the Submission Abstract



Note. The upper tail of writing scores is mostly all human-written text, both pre- and post-Chat GPT.

Figure 8 shows coefficient sizes for the relationship between AI Scores for abstracts and various writing quality measures. All models include fixed effects for time (quarters) and editor; all standard errors are two-way clustered at these levels. Online Appendix A4 contains the definitions of each of our writing quality measures. We find strong evidence that AI use is associated

Figure 6. Trends in Flesch Reading Ease (Abstracts) from 2013 to February 2026 with Markers for COVID-19 and the Launch of ChatGPT

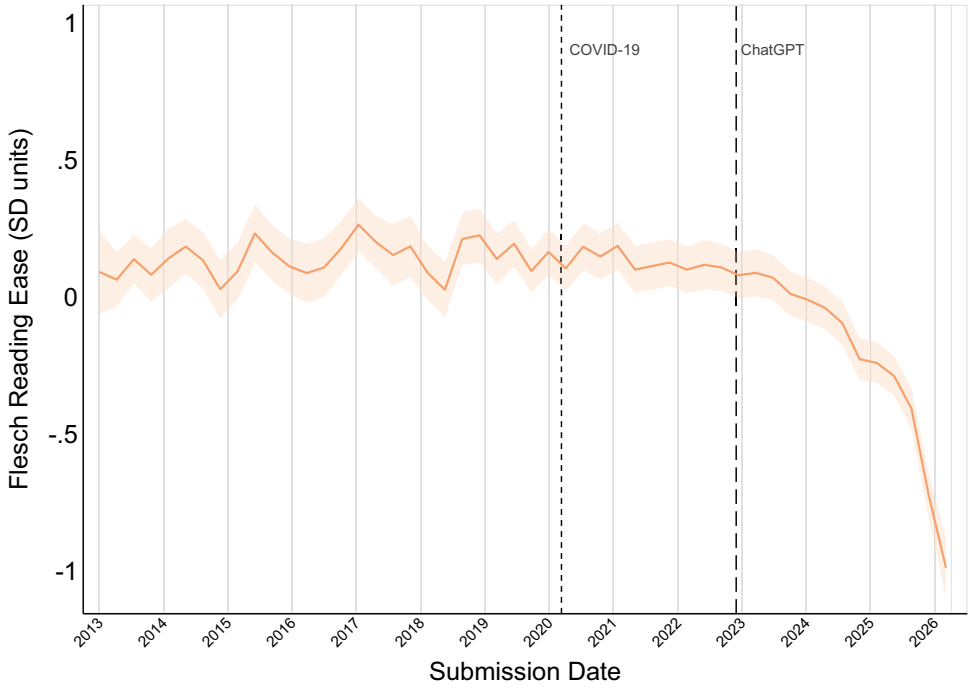
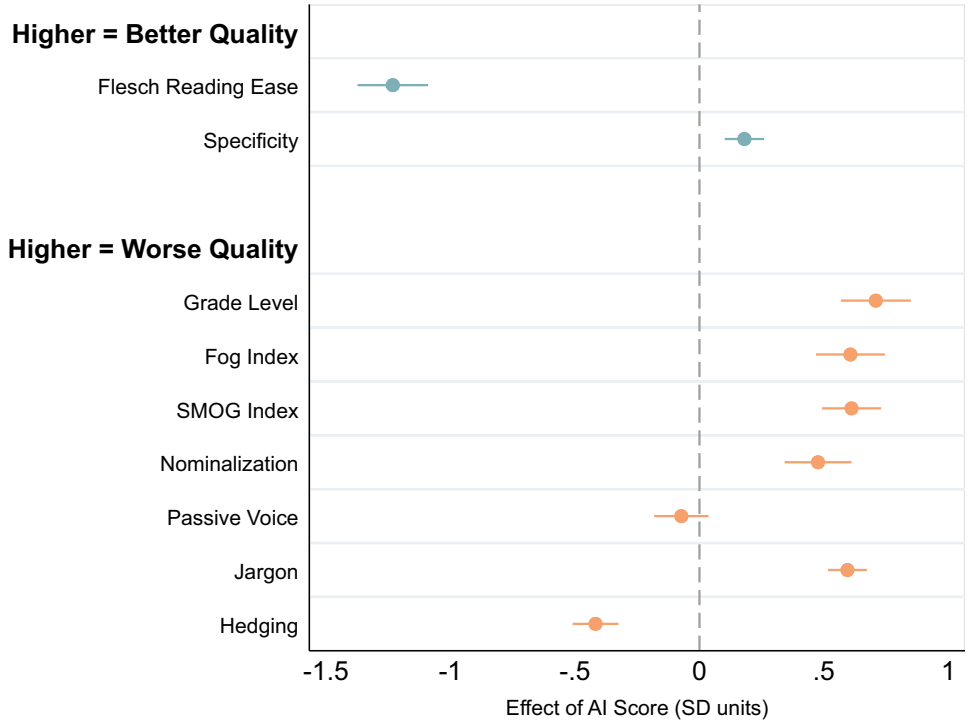


Figure 8. Association Between AI Use Categories and Different Measures of Writing Quality



Note. See Online Appendix A4 for definitions of each writing measure and how it is calculated.

with lower-quality writing across most of these traditional measures. This result is counterintuitive. Authors often assume that using AI will improve their writing. However, this is not the case, at least when authors substantially offload their writing to it.

AI prose is more difficult to read on several dimensions. Beyond substantially lower Flesch Reading Ease scores, the grade level required to understand the text is higher (more multisyllabic words); the FOG and SMOG indices increase, suggesting more complex text; and the use of jargon increases. We also find increased use of nominalizations (e.g., "conceptualization", "operationalization", or "contextualization"). We do see better scores in some writing measures, with less passive text, less hedging, and greater specificity (e.g., mentioning numbers).[7] Overall, AI text is more direct and specific, but also more difficult to read.

3.4. AI Use and Editorial Outcomes

Writing metrics are only one aspect of quality. Manuscripts also have intellectual content. However, evaluating a paper's research content is much more challenging to do algorithmically. Fortunately, in our current academic system, journals are tasked with exactly this effort. At *Organization Science*, the review process unfolds as follows. First, a managing editor evaluates the manuscript for basic flags, including plagiarism or any other indicator that the manuscript should not be forwarded to a deputy editor for a decision. Manuscripts that pass the initial plagiarism check

Table 1. Editorial Pipeline by AI Category (Post-ChatGPT)

	0%–15% AI	15%–30% AI	30%–70% AI	70%+ AI
Total Manuscripts	2,754	771	717	473
Desk Rejected	1,204 (43.7%)	300 (38.9%)	407 (56.8%)	329 (69.6%)
Sent to Review	1,550 (56.3%)	471 (61.1%)	310 (43.2%)	144 (30.4%)
Rejected After Review	1,124 (72.5%)	312 (66.2%)	222 (71.6%)	83 (57.6%)
Received R&R	329 (11.9%)	101 (13.1%)	40 (5.6%)	15 (3.2%)
Overall Rejection Rate	84.6%	79.4%	87.7%	87.1%
Mean Readability (std)	0.04	−0.33	−0.61	−1.27
Mean Readability, std (Survivors)	0.10	−0.30	−0.52	−1.09

Notes. Sample: First submissions after the launch of ChatGPT (November 30, 2022). Desk Rejected and Sent to Review percentages are of Total Manuscripts. Rejected After Review percentage is of those Sent to Review. R&R percentage is of Total Manuscripts.

Table 2. Predictors of Desk Rejection

	Desk Reject	Desk Reject	Desk Reject	Desk Reject	Desk Reject
Readability (std)	−0.027***	−0.028***	−0.023***	−0.023***	−0.023***
	(0.008)	(0.008)	(0.007)	(0.007)	(0.007)
AI Score	0.299***	0.266***	0.187***	0.188***	0.191***
	(0.035)	(0.025)	(0.020)	(0.021)	(0.025)
All Non-Native English			0.322***	0.321***	0.322***
			(0.011)	(0.012)	(0.013)
AI Score × Non-Native English					−0.006
					(0.044)
_cons	0.385***	0.390***	0.282***	0.282***	0.282***
	(0.022)	(0.008)	(0.008)	(0.005)	(0.005)
Observations	6,725	6,725	6,725	6,724	6,724
R^2	0.034	0.058	0.154	0.163	0.163
Quarter FE		Yes	Yes	Yes	Yes
Editor FE				Yes	Yes

Notes. Standard errors in parentheses. All models use two-way clustered SE by quarter and editor.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

are sent to a deputy editor, who decides whether the paper should go out for review and, if so, which senior editor should handle it. If a manuscript moves forward, a senior editor decides whether to send the paper out for review or desk-reject it, with these assignments more closely aligned with the senior editor's expertise than those of the deputy editor, who operates more as a generalist within a subfield. Finally, reviewers provide written feedback and a recommended decision, followed by the editor's final determination. In short, the editorial and peer-review process is intensive and expressly designed to assess papers on their intellectual content.

Table 1 provides descriptive evidence of manuscript outcomes by level of AI use, restricted to the post-ChatGPT period to ensure comparability. A clear pattern emerges above 30% AI use. Manuscripts crossing this threshold have desk rejection rates nearly 30 percentage points higher than those below it, a 60%–80% relative increase. By the end of the editorial funnel, only 3.2% of high-AI papers (70%+ AI) receive Revise & Resubmit decisions, compared with over three times that rate among low-AI submissions. In other words, editors, likely without recognizing the writing as AI-generated, consistently judged these manuscripts as lower quality and unworthy of reviewers' time. Interestingly, submissions in the 15%–30% range show a slight increase in success rates, possibly reflecting a more human-first collaboration mode or differences in authors, topics, and other factors that we cannot disentangle. Nevertheless, this slightly elevated R&R rate from 0%–15% (11.9%) to 15%–30% (13.1%) is not statistically different ($p = 0.387$). As such, the 30% threshold marks a sharp divide, with manuscript prospects declining steadily above it.

In Tables 2 and 3, we examine the relationship between AI use and editorial decisions at the initial screen and after review, respectively. These models also include quarter and editor fixed effects, with standard errors cross-clustered at these two levels. Table 2 shows that higher writing quality is associated with

Table 3. Predictors of a Paper Being Rejected After It Is Sent Out for Review

	(1) Reject	(2) Reject	(3) Reject	(4) Reject
AI Score	0.086***	0.075***	0.072***	0.108***
	(0.023)	(0.024)	(0.023)	(0.032)
All Non-Native English		0.082***	0.081***	0.099***
		(0.015)	(0.015)	(0.014)
AI Score × Non-Native English				−0.143**
				(0.066)
_cons	0.715***	0.697***	0.698***	0.694***
	(0.002)	(0.003)	(0.002)	(0.002)
Observations	3,801	3,801	3,800	3,800
R^2	0.148	0.154	0.160	0.161
Editor FE			Yes	Yes

Notes. Standard errors in parentheses. All models include quarter FE and two-way clustered SE. Sample: Non-desk-rejected submissions only.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

lower rates of desk rejection. More importantly, manuscripts with more AI also have a significantly higher rejection rate (+0.293 above the baseline of 0.399), independent of writing quality. This suggests that the higher rejection rates are not due only to poor writing, but also to underlying quality issues that we cannot observe in this analysis, but editors and reviewers do when assessing each paper on its own.

We next test whether AI use by *All Non-Native English* teams helps them in the review process. Many authors use AI with the assumption that it will improve their writing, thereby making publication easier. As a baseline, we find that *All Non-Native English* teams do indeed have an elevated desk rejection rate. The high rate may be due to several factors: topic relevance, theoretical framing, methods, differences in writing style, among others. Including this term decreases the "AI score" coefficient from 0.266 to 0.187. Nevertheless, these appear to be two distinct negative correlations in the review process. In column (5), we interact the AI score and the *All Non-Native English* variable. We find little evidence that AI use by authors from nonnative English-speaking countries is related to more favorable decisions. The coefficient on the interaction term is near zero. These effects persist even after accounting for editor fixed effects and time trends.

The specific mechanisms behind these patterns are beyond the scope of this analysis. It is worth noting that, when the author fixed effects are included in Table A5, for example, using manuscripts by the same author submitted with high and low-AI use, the effect of AI becomes insignificant for manuscript decisions, but the effect of writing quality remains significant and negative as can be seen in Table A4. Nevertheless, a key takeaway from this analysis is that AI writing does not appear to benefit authors.[8]

In Table 3, we analyze whether AI use affects the editorial outcomes of papers sent out for review by a senior editor. The pattern of results is similar: greater AI use is associated with higher rejection rates. We again find that *All Non-Native English* teams have higher rejection rates. However, conditional on overcoming the desk rejection threshold, the two coefficients—of AI *and* non-Native English—on rejections are reduced by nearly half. That is not to say that non-Native English authors are helped by AI at this stage, but rather that they are not doubly hurt because they use AI. Nevertheless, the main action for heavily AI-driven manuscripts is at the desk-reject stage.

Together, these findings suggest that AI use in submissions to *Organization Science* is of the "more" rather than "better" variety. The marginal AI submission is of lower quality and is more likely to be desk-rejected. AI usage appears to reflect underlying information about the quality of the manuscript, information that is being picked up in the review process by both deputy editors and senior editors—leading to higher rates of desk rejection and rejection after review. Furthermore, given the time period we analyze, it is unlikely that editors flagged these manuscripts as "AI." Rather, a general quality assessment was conducted, and high-AI papers were deemed substantially worse, often not worth the reviewers' time.

3.5. Are Publication Incentives Driving the AI Surge?

The rise in low-quality AI-augmented submissions raises the question: *What incentives are driving this behavior?* While academics might write more low-quality papers for many reasons, one explanation is a strategic response to supply-side incentives. In recent years, many business schools have adopted "journal lists"—a curated set of journals that *count* toward tenure and promotion. Some schools award significant financial rewards to faculty publishing in these specific outlets. A prominent example of a widely used journal list is compiled by UT-Dallas (hereafter, UTD) and includes 24 journals, including *Organization Science*. Journals on this list are used to rank the top 100 schools by publication output. Many schools also "count" publications based on the FT50 list or the Academic Journal Guide system (or their own custom variants), but the UTD journals are almost universally in the top category in all of these.

Institutions do not publicly announce their publication incentives; therefore, we try to infer this from differences in publication patterns across schools. When the UTD rankings were introduced in 2005, we observed a sharp discontinuity in the publication response of some schools. To identify "UTD Responder" schools, we compare each school's average annual UTD publications before (1990–2004) and after (2005–2020) the rankings were introduced. We classify schools whose *difference* between pre- and post-UTD publication volume is *above the median* as Responders. We then match each responder to a nonresponder based on having an equivalent pre-UTD (e.g., prior to 2005) publication record. Critically, the designation of a school as a UTD Responder is based on data generated *prior* to ChatGPT's launch in November 2022. In our matched sample analysis below, we have 48 schools that serve as our "control" non-Responder schools, and 48 matched schools designated as "UTD Responders."

We use this data to test whether supply-side incentives drove AI adoption. We estimate the following specification at the school × month level:

$$Y_{s,t} = \alpha_s + \gamma_t + \beta(\text{UTD Responder}_s \times \text{Post-ChatGPT}_t) + \varepsilon_{s,t} \quad (1)$$

where α_s and γ_t denote school and month-year fixed effects, respectively. The school fixed effects account

for time-invariant differences in business school submission patterns to *Organization Science*, while the month-year fixed effects absorb common time shocks such as overall submission trends (e.g., the journal publishing more papers). UTD Responder$_s$ equals one if the school is classified as an above-median responder, and Post-ChatGPT$_t$ equals one for months from November 2022 onward. Standard errors are two-way clustered by school and month-year. Note that schools are included as matched pairs with equivalent pre-UTD (pre-2005) publication volumes but divergent volumes afterward (until 2020). It is important to note that the "UTD Responder" designation is fixed before our sample period starts in January 2021.

Our dependent variables Y are calculated as the number of monthly submissions each school makes to *Organization Science*. Specifically, we have five dependent variables: total submission volume, and submission counts decomposed into four AI-intensity bands—0%–15%, 15%–30%, 30%–70%, and 70%+ AI score—allowing us to identify which types of submissions are driving any aggregate effect.

In Table 4, we test this supply-side incentive mechanism. We focus our analysis on two sample populations. In Panel A, we include all schools in our analysis. We find that the coefficient of interest (UTD Responder × Post-ChatGPT) is positive and significant for overall volume, and all AI categories above 15%. UTD Responders, after ChatGPT's launch, increased their AI submissions substantially more than schools without "counting" incentives. In Panel B, we exclude schools based in Mainland China and Hong Kong from our analysis. This exclusion is intended to address concerns that our results are driven by a subset of schools with a fastest growing researcher population and submission counts. The results in Panel B are directionally similar.

In short, AI adoption—in particular, heavy reliance on AI writing in manuscripts—is not solely a function of the introduction of language models themselves. Instead, it appears to reflect a rational response by authors to their home institution's incentives, particularly within schools that emphasize journal lists and publication counts. We caution, however, that our measure of school incentives is noisy. This topic deserves further, more rigorous analysis.

4. AI in the Review Process

4.1. AI Use in Reviews over Time

We now turn our attention to the review side of the process. We again apply Pangram AI detection to the submitted reviews in *Organization Science*. We analyze reviews entered into the system as raw text, which accounts for 79% of the total.

We first analyze the proportion of reviews categorized in our AI-intensity bands—0%–15%, 15%–30%, 30%–70%, and 70%+ AI score. While AI use in reviews is not as pervasive as in manuscripts, over 30% of reviews in *Organization Science* still use some degree of AI today. The fastest growing category consists of reviews classified as 30%–70% AI, a category that our experiments suggest consists of substantial use of AI in the writing process. These results are presented in Figure 9.

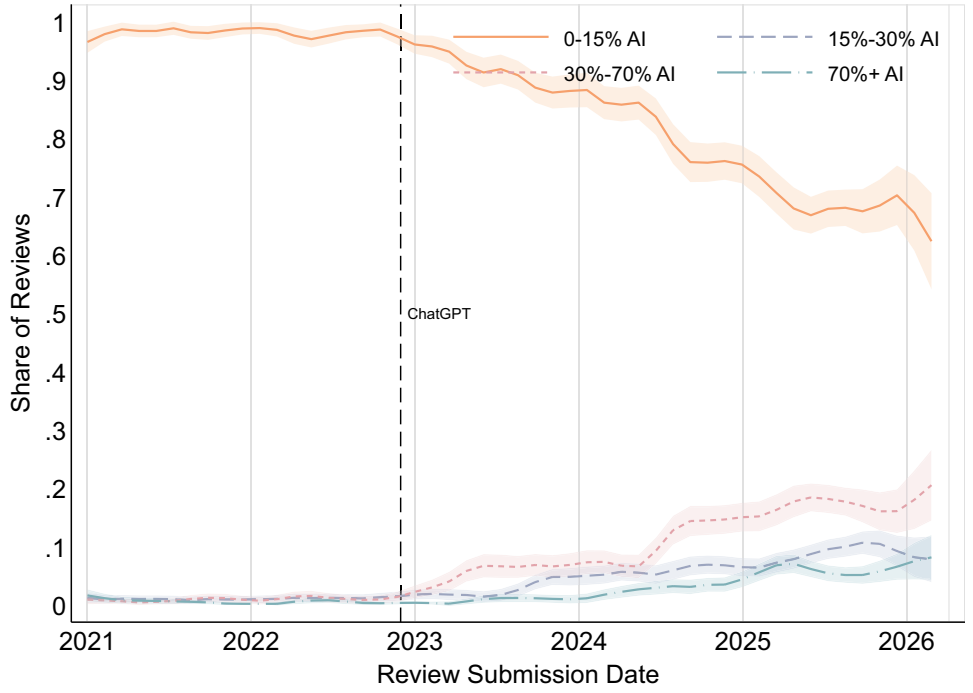
Next, in Figure 10, we report on the quality of writing in these reviews over time, standardized to a mean of zero and a standard deviation of one. As with paper submissions, we observe a similar, marked decline in the Flesch Reading Ease of the reviews

Table 4. Do Schools That Responded to UTD Rankings Also Respond to the Availability of AI?

	(1) Volume	(2) 0%–15% AI	(3) 15%–30% AI	(4) 30%–70% AI	(5) 70%+ AI
Panel A: All schools					
UTD Responder × Post-ChatGPT	0.101**	−0.005	0.060***	0.032***	0.014*
	(0.040)	(0.029)	(0.016)	(0.010)	(0.008)
Observations	6,048	6,048	6,048	6,048	6,048
R^2	0.250	0.185	0.109	0.076	0.074
School FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
Panel B: Excluding Chinese Mainland and Hong Kong					
UTD Responder × Post-ChatGPT	0.067*	−0.009	0.048***	0.024**	0.007
	(0.036)	(0.027)	(0.015)	(0.009)	(0.006)
Observations	5,796	5,796	5,796	5,796	5,796
R^2	0.231	0.171	0.103	0.071	0.061
School FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes

Notes. Two-way clustered SE by school and month-year. 1:1 nearest-neighbor matched on pre-2005 publication levels (no replacement). Classification uses 1990–2020 data only (pre-ChatGPT). 0%–15% AI: <0.15; 15%–30% AI: 0.15–0.30; 30%–70% AI: 0.3–0.7; 70%+ AI: >0.7.

Figure 9. AI Use in Reviews Over Time



Notes. The figure shows the share of reviews classified as AI across different thresholds from 2021 through early 2026. The dashed vertical line marks the launch of ChatGPT in November 2022.

submitted to *Organization Science* following the launch of ChatGPT.

Similarly, we find a concerning pattern of degradation in the written reviews across nearly all of our writing quality measures, as shown in Figure 11. Reviews that use AI are harder to read, with more complex word choice, more nominalization, and more jargon.

Figure 10. Trends in Flesch Reading Ease (Reviews) from 2013 to February 2026 with Markers for COVID-19 and the Launch of ChatGPT

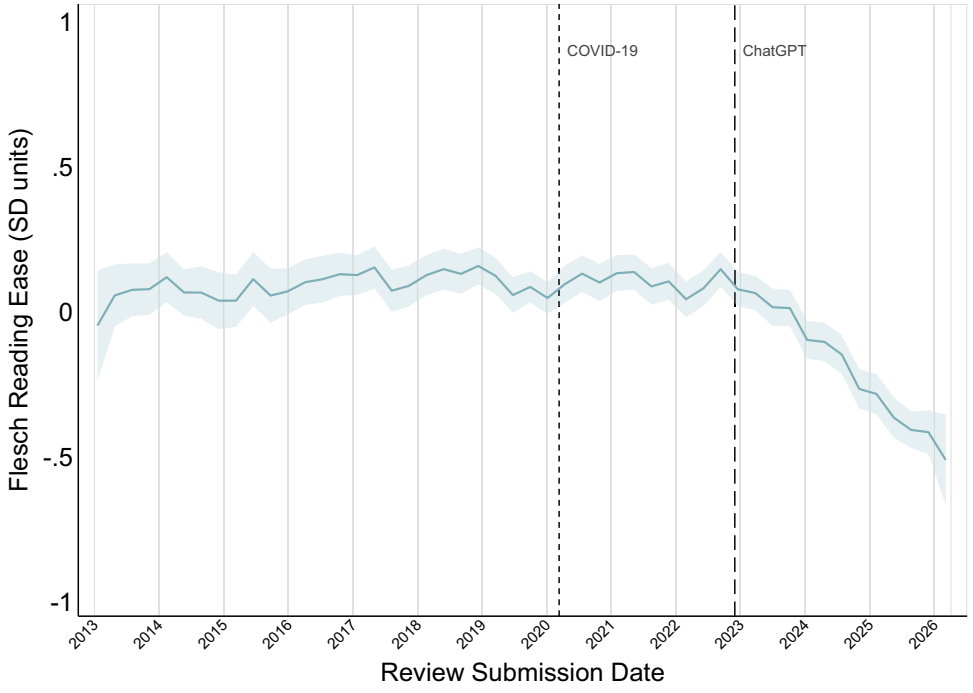
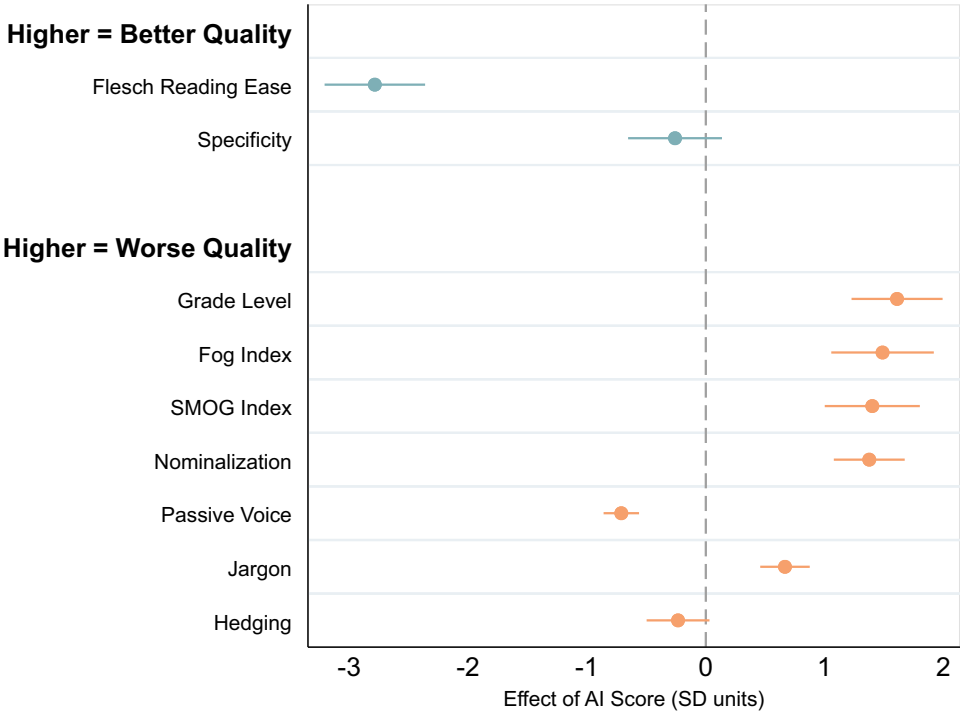


Figure 11. Association Between AI Use in Reviews and Review Writing Quality



Notes. The figure displays regression coefficients from a model predicting a standardized writing quality index as a function of AI use category indicators, with quarter and editor fixed effects. Online Appendix A4 contains the definitions of each measure and descriptions of how they are calculated.

Overall, these trends suggest that the growth of AI reviews will make it harder for both editors and authors to act upon reviewer feedback. This increases the work of editors, who now have to spend more time understanding reviews, and of authors, who must interpret and respond to reviewers' requests and recommendations.

4.2. AI Use and Topics Raised in Reviews

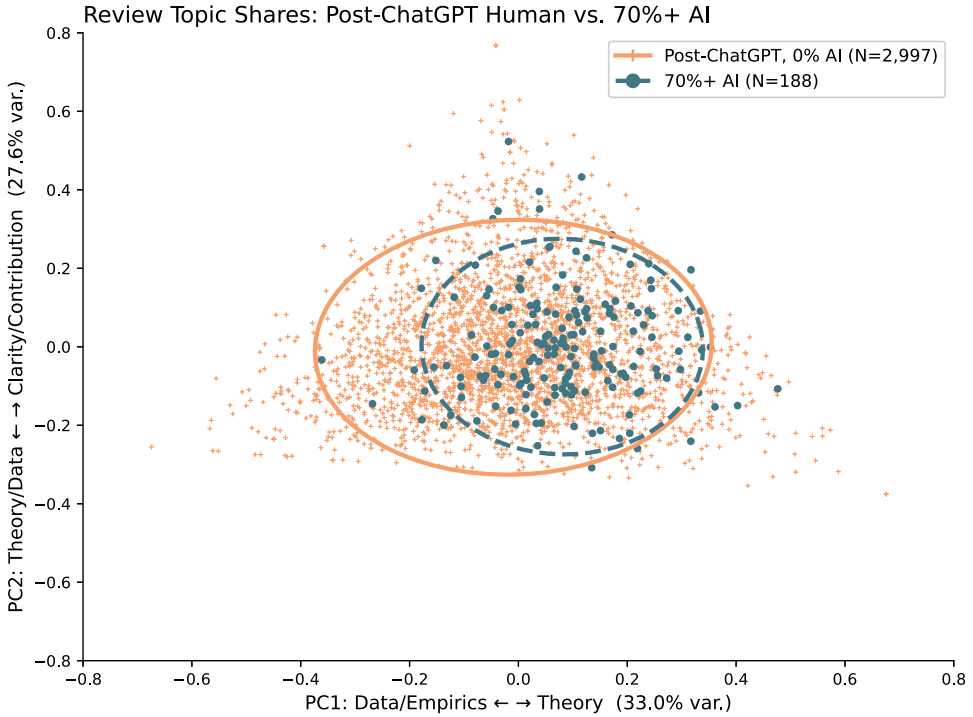
A decline in writing quality is only one dimension where AI might affect the review process. Reviews also provide authors with feedback on specific aspects of their manuscripts that are weak or need further development. Reviewers often focus on five areas: theory, contribution, clarity of the argument and writing, data, and empirics. In Table 5, we regress each topic's relative share of attention in a review as a function of that review's AI score. These models also include fixed effects for the manuscript. This allows us to compare two reviews for the *same* manuscript while holding constant differences in the nature and content of a specific research paper. Furthermore, we include reviewer fixed effects. Different reviewers may emphasize different aspects of the manuscript, potentially biasing the relationship between topic emphasis and AI score. This specification gives us an ideal basis for comparison, as we can partial out both of these factors. We find strong shifts in emphasis even with these fixed

Table 5. Are AI Reviews Different than Human Reviews? (Reviewer Fixed Effects)

	(1) Theory	(2) Contribution	(3) Clarity	(4) Data	(5) Empirics	(6) Readability	(7) Word Count
AI Score	0.251**	0.127	−0.016	−0.279**	−0.102	−2.388***	26.623
	(0.106)	(0.116)	(0.104)	(0.109)	(0.108)	(0.111)	(48.236)
Constant	0.005	−0.023***	−0.021***	0.050***	−0.014*	0.108***	1119.770***
	(0.008)	(0.008)	(0.007)	(0.008)	(0.008)	(0.008)	(3.511)
Observations	7,208	7,208	7,208	7,208	7,208	7,570	7,570
R^2	0.743	0.727	0.741	0.741	0.737	0.839	0.806
Manuscript FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Reviewer FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes. Standard errors in parentheses. Cols 1–5: DV is topic share normalized to SD from pre-ChatGPT mean. Col 6: Flesch Reading Ease (standardized, mean 0, SD 1; higher = more readable). Col 7: Review word count. Standard errors clustered at the manuscript level.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

Figure 12. Are AI Reviews Different from Human Reviews?



Notes. This represents a narrowing of the types of topics these reviews cover. Each data point represents one review. The *x*-axis plots the first principal component and the *y*-axis plots the second principal component. See text for description of how the principal component analysis was conducted.

effects. AI use biases reviews toward a "theory" focus and away from a "data" focus. Moreover, AI-generated reviews are of lower writing quality, even after controlling for differences among reviewers. AI is substantively changing the content of the reviews, shifting the focus of editors and authors, and potentially affecting manuscript quality as papers are revised in response to these narrower concerns.

Finally, Figure 12 provides a visual depiction of this narrowing of review reports. To produce this figure, we conduct Principal Component Analysis (PCA) on our five review topic categories and then plot each review as a point on the graph. The *x*-axis shows a point's location on the PC1 axis representing the Data/Empirics versus Theory distinction. Points to the right represent reviews with a greater emphasis on Theory, whereas points to the left represent reviews with a greater emphasis on Data/Empirics. The *y*-axis, PC2, represents the Theory/Data (substance) versus Clarity/Contribution (presentation) axis. Reviews higher on the *y*-axis focus more on presentation; lower on substance.

The plus signs (orange) are reviews with 0% AI detected after the launch of ChatGPT; the dots (teal) ones are 70%+ AI reviews during this same period. Each ellipse encompasses the 95% boundary for each review type. When AI is used for reviews, these reviews cover a narrower evaluative range. That is, review diversity systematically decreases. The results are exactly the same if we limit the number of 0% AI reviews to a random $n = 118$, as can be seen in Online Appendix Figure A1.

Table 6. Do AI Reviews Increase the Likelihood That a Paper Is Rejected?

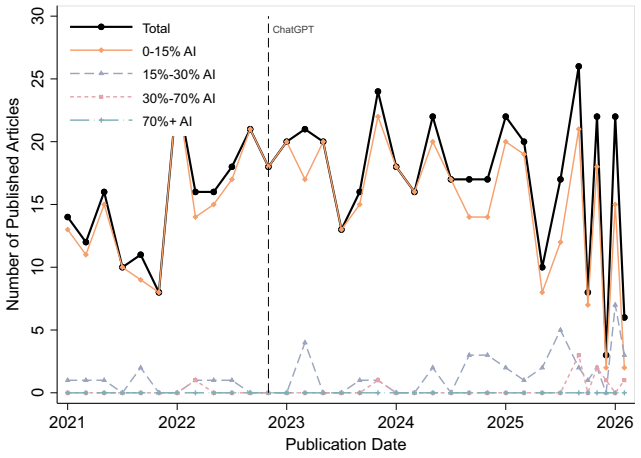
	(1) Reject	(2) Reject	(3) Reject
Mean Review AI Score	0.033	0.037	0.046
	(0.061)	(0.061)	(0.065)
Manuscript AI Score	0.109***	0.103***	0.113***
	(0.027)	(0.024)	(0.035)
Constant	0.746***	0.746***	0.744***
	(0.004)	(0.003)	(0.002)
Observations	3,470	3,469	3,461
R^2	0.061	0.067	0.120
EIC FE		Yes	Yes
Senior Editor FE			Yes

Notes. Standard errors in parentheses. All models include quarter FE and clustered SE. Sample: Non-desk-rejected submissions only.
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

4.3. AI Reviews and Editorial Decisions

We next test whether AI reviews are systematically shifting editorial decision-making. In Table 6, we find little evidence that having AI-supported reviews is affecting the decision to reject a manuscript or move it forward in the process. It is very likely that editors are ignoring these reviews and substituting them with

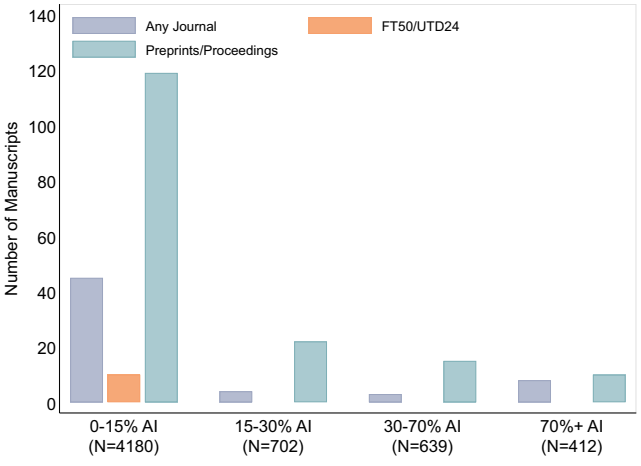
Figure 13. Trends in AI Usage Among Articles Published in *Organization Science*



Notes. Each line shows the share of published article abstracts classified as 0%–15% AI, 15%–30%, 30%–70%, 70%+. The dashed vertical line marks the launch of ChatGPT in November 2022.

their own judgment or relying more on the comments of the other, human, reviewers.[9] Human expert judgment is still limiting any adverse effects of AI on what we publish, but is doing so at the cost of increased effort. Figure A8 shows the average number of Deputy Editors, Senior Editors, and Reviewers handling manuscripts over time from 2021 to 2026. To address the growing volume of submissions, the journal increased the number of Deputy editors from six to eleven. The number of active senior editors has also increased from approximately 30 in the preperiod to 60 in the postperiod. This large jump in effort by volunteer editors has allowed Organization Science to maintain (for DEs and reviewers) or, in some cases, decrease the active editorial load (for SEs)—even though our submission volume has increased by over 40%. A much larger number of DEs are handling a growing number of submissions that need to be desk-rejected, with some DEs having annual loads of over 250 manuscripts.[10]

Figure 14. Distribution of Publication Outcomes for Articles Rejected at *Organization Science*, by AI Score



Note. Preprint servers and Proceedings include SSRN, arXiv, as well as the Academy of Management Proceedings.

4.4. What Appears in Press?

While it is too early to tell how the recent rise in AI submissions will translate into publications, we examined how many of *Organization Science*'s eventual publications contain detectable levels of AI. We downloaded all Organization Science publications in the OpenAlex database and scored these using Pangram. These results are presented in Figure 13. The vast majority of published papers are human-generated (<15% AI score), with a few showing AI collaboration (15%–30% score) in the writing of the abstract. This is reassuring, but it also suggests that the editorial process puts considerable effort into weeding out low-quality papers, many of which rely heavily on AI.

Finally, in Figure 14, we tracked the outcomes of manuscripts rejected at *Organization Science*. The vast majority of these papers could not be found on preprint servers, in proceedings, or in other academic journals, likely because many of the rejected articles are still under review elsewhere. However, the pattern of results among those we were able to track is informative. Manuscripts with very low AI scores (0%–15%) were most likely to be published in any journal (many of which are well-respected journals), and some were published in the very best journals in our field. Publication rates are substantially lower for articles with AI scores between 15% and 70%. What is surprising is that, while no articles with an AI score of 70% were eventually found in "top" or otherwise selective journals, these manuscripts are the most likely to appear in print at any journal, at a rate of 1.9% (versus 0.8% of papers with an AI score of <15% AI. This pattern suggests that heavily AI papers can find publication homes at peer-reviewed outlets.

5. What Comes Next?

Our evidence shows how AI-based writing is flooding *Organization Science* with a high volume of low-quality papers. Institutional incentives—particularly when careers depend on publication counts in specific journal lists—appear to be accelerating this trend. This surge has strained the volunteer labor that evaluates research, which might partly explain the high percentage of reviews with AI-generated signatures we observe. Like AI-generated manuscripts, these AI reviews are less well-written and narrower in scope. In short, AI is placing the peer-review system under stress that shows no signs of decreasing. Although ours is the first comprehensive study of a journal's

full submissions corpus, it is hard to imagine that *Organization Science* is alone in this regard. Our conversations with editors at other journals reveal anecdotal accounts of these same patterns; these are simply the first data to confirm them.

Importantly, our findings temper some of the optimistic predictions about how AI changes academic research by showing:

- AI is not improving writing quality in aggregate. Our evidence suggests that the written quality of manuscripts and reviews is in decline. This is not to say that AI *cannot* improve writing quality, but so far, the systems-level effects have been negative.
- AI is not leveling the playing field. All non-Native English Institution author teams and new field entrants who expect AI to strengthen writing receive little benefit in terms of increased publication likelihood.
- AI reviews are not equivalent to (or better than) human reviews. They are less well-written and narrower in systematic ways.
- AI adoption is not a purely technical phenomenon or institutionally agnostic. Instead, it appears to be correlated with career incentives that pressure researchers to increase their publication output.

5.1. Does AI Use Necessarily Make Writing Worse?

Before considering how institutions might respond, we should clarify our position on the deterioration in writing quality.[11] We do not see the deterioration in writing quality that we document as an inevitable consequence of AI use. The worst outcomes likely reflect how authors and reviewers use the tools today—as a substitute for their writing and human editing rather than a complement.

Indeed, we used AI in preparing this essay. On the back end, it helped write the code to generate the figures and tables. In writing, we used AI to help us outline sections, find better phrasing, and compare our essay to prior research, among various other uses. The manual revisions and copyediting by our author team and other human readers, however, were extensive.

Even with considerable AI use in its production, this editorial received a Pangram score of 8.8%.[12] It was classified as "100% Human" by both Pangram and our own 15% or 30% cut-offs for human-first papers. Given our own use of AI, it has forced us to reflect on what higher scores, for instance 30% or more, say about a manuscript's production. Submitted manuscripts with scores below 30% receive editorial outcomes that are indistinguishable from those with a score near 0%. Above it, outcomes deteriorate.

At least two mechanisms may be driving this. The first is strategic: authors with weaker manuscripts may turn to AI as a last resort, hoping that polished prose can compensate for actual time spent working on the manuscript. The second may be what Shaw and Nave (2026) term "cognitive surrender," the tendency to adopt AI outputs with minimal scrutiny, bypassing the deliberative reasoning that careful academic writing demands.[13]

These mechanisms have different implications. For those tempted by the strategic approach, the data are clear: heavily AI-generated submissions are usually rejected. And for all of us, it is critical to evaluate *when* engaging with AI sharpens thinking and when it replaces it.

5.2. Toward a Better Equilibrium

These individual choices matter, but they alone will not resolve the systemic pressures driving AI adoption. Steering toward a "better" versus "more" equilibrium requires institutional change.

Moving toward a "better" equilibrium requires reframing of AI's role in academic research. Most authors and reviewers are using AI to compress time, perhaps unaware of the degradation in quality. A paper can now be written in days, not weeks or months. Reviews can be submitted in minutes instead of hours. The same dynamic extends to empirical analysis, where AI-guided coding accelerates data cleaning and estimation but may erode the nuance and judgment that careful empirical work demands. Future AI models may reduce this concern, but our analysis shows the importance of measuring this progress rather than assuming it. The writing quality needed to publish in strong journals such as *Organization Science* is currently beyond what Dell'Acqua et al. (2026) call "the jagged frontier" of even the latest AI. Tracking this frontier is essential, both to address the apparent overconfidence scholars display in their AI-generated content (Leonardi and Leavell 2026) and to guard against applying algorithmic tools without understanding how or how well they work (Anthony 2021).

We also recognize that while we measure one dimension of research *outcome* quality, many elements of *process* quality remain important for research progress but are not directly observable in our data. Many scholars consider research to be a craft: a lifelong practice of learning to engage the world, interacting with other scholars, reading deeply, and refining our cognitive schemas through the work itself (Bechky and Davis 2025). This craft develops through the struggle of doing, whether that means iterating on an empirical strategy, immersing in a field site, or wrestling a theory into coherent form. If researchers systematically delegate these processes to AI, we risk eroding the very capacities that enable us to recognize meaningful empirical puzzles, draw novel connections, and generate collective knowledge at the frontier.

5.3. Peer Review: Potential Scenarios

The institution of peer review is younger than we generally assume. Although informal refereeing dates back centuries, the modern system of anonymous peer review did not become standard until well into the second half of the twentieth century. *Nature* did not require external peer review for all submitted articles until 1973; *The Lancet* followed in 1976. The system that emerged was built on a specific social contract: anonymous volunteers would donate their expertise for the good of the field, in exchange for the trust that their judgments would shape what counted as legitimate science.

That system has evolved several times, from editorial discretion to formal refereeing to the institutionalized peer review we recognize today, and it will surely evolve again—the question is toward what. Whether we get to *better* or stay on a trajectory toward *more* is not clear. We present three possible scenarios for how the system may evolve, based on what we observe today.

In **scenario one**, our findings could represent a transitory phase. AI-generated papers are faring poorly in the review process, and it is very likely that, as detection improves, even more of this type of work will be filtered out of the system. But if the deeper cause is strategic institutional behavior and the incentives are deeply ingrained, we may be facing more of the same for some time.

Scenario two is a further erosion of the field as the trends we report here continue. Volume will eventually overwhelm the best reviewers and editors, who will understandably withdraw from increasingly exhausting volunteer work. This might leave AI review agents evaluating authors' AI-generated submissions, hollowing out the shared intellectual community we know advances science. This may be the most likely path if conservative academic journals and universities fail to address this potential outcome.

A **scenario three** also exists, however, in which institutional changes and technological improvements in artificial intelligence lead to a "better" equilibrium where AI enables better research questions, theory, and data analysis. Rather than evolving inevitably, this better equilibrium entails deliberate choices that extend well beyond journals. It is the hardest path and the one most worth steering toward. That steering requires changing what the two institutions that most shape scholarly incentives—journals and universities—choose to reward.

5.4. What Does Better Look Like?

Before considering how institutions should adjust what they reward, it is important to consider what "better" research may look like. It is hard to predict exactly how better research in collaboration with AI will evolve. This will depend heavily on researchers' ingenuity and the evolving capabilities of AI models. But we see three general categories of work emerging.

The first is our ability to apply better methods to existing questions. This includes using AI in creative and new ways to tackle critical questions in our field for which we historically had limited tools or data. In this way, AI is like the many historical and emerging technologies and data sources we've already incorporated into our research. AI could help us better analyze large-scale text, video, audio, and images at a scale previously infeasible.

The second is to use AI to identify new and important questions that were not visible because they existed in the gaps between research fields. AI, with its ability to find connections in seemingly disparate domains and fields—across data forms such as images and text and across languages—may reveal those gaps in theory and evidence more readily than any human would. The potential for AI to bridge knowledge silos is particularly exciting for our diverse field (Dell'Acqua et al. 2026). Humans may continue to hold the key to answering these questions, possibly with conventional methods; it could be the machine's ability to span vast amounts of knowledge that creates the opportunity.

Finally, the most ambitious category consists of both new questions and new methods. As researchers experiment with AI, they will discover things that are unfamiliar and imperfect by conventional standards. While these efforts are risky and time-consuming, they can also change the field.

5.4.1. Making the System Sustainable and the Research Stronger.

The most immediate operational challenge is the rise in volume. Our evidence implies that the sharp rise in submission rates at *Organization Science* since late 2022 was driven by AI. If this trajectory continues, the volunteer-based review system will be unable to keep pace. This poses an important organizational design challenge (Hasan et al. 2025). Journals need tools and policies that maintain volume at sustainable levels or facilitate low-cost rejections of low-quality AI-driven research. We emphasize that the goal is not to screen out research enabled by AI, since our objective is to refine and promote great research on organizations. Artificial intelligence provides powerful tools for conducting research, many of which we as authors use to support our own work. Our data show that AI use in aggregate is not currently producing high-quality writing, but that does not mean it cannot in the right hands or with further advances in technology.

AI detection tools such as Pangram can also play a role here, used as additional flags that help editors understand how a manuscript was produced and evaluate it in context. We do not advocate for detection as a gatekeeper—automated desk-reject decisions based on AI scores would be premature and invite

gaming behavior—and we are uncomfortable making normative assessments about appropriate levels of AI in submissions. But detection data can inform editorial triage, flagging submissions that warrant closer scrutiny before reviewers are asked to invest their time. It will be crucial for studies like ours to be frequently repeated as technology advances and researchers' behaviors adapt. The AI-driven writing we are detecting through Pangram in March 2026 sends a strong signal of lower quality, but that signal is noisy and may change if leading-edge models advance their writing skills.

Journals should also consider whether appropriately designed submission fees can help internalize the costs that high-volume submitters impose on the review system. A nonlinear fee structure that escalates costs for authors with simultaneous submissions could specifically target the volume incentive that AI-generated manuscripts exploit, while imposing minimal burden on authors submitting at typical rates. Such fees would need to incorporate equity considerations for those with limited resources, and we are also wary of taxing researchers who are truly producing a volume of high-quality work. We raise these options as possibilities worth serious discussion, not as settled recommendations.

Reducing submission volume is necessary but insufficient if the research that does get published is not improving. Editors and reviewers must explicitly signal that they are looking for work that pushes the frontier, even when that work is messier or less immediately comparable to established approaches.

Universities. The imbalance between quantity and quality cannot be fully addressed by journals themselves if universities continue to primarily reward volume in career progression. Volume incentives originate not at the journals but at the universities themselves, and even more so at those that reward simple publication counts and offer financial bonuses for placement in designated "A" journals. While these incentives have always been distortionary, the falling cost of producing submissions will exacerbate the problem. At a minimum, institutions should be transparent about these incentive structures. Disclosure of publication bonuses and explicit journal-list-based rewards would allow the field to see the scope of the problem and begin to design responses.

This requires institutional changes at each major stage of the academic career ladder. Tenure and promotion decisions carry the highest stakes and demand the most significant shift. At a minimum, existing norms and rules regarding the exact count of publications required to define someone as sufficiently "productive" will be outdated; the more fundamental change will require rewarding exceptional quality over quantity. Committees must do the hard work of actually reading papers and assessing intellectual contribution rather than counting publications. While this has always been the aspiration, it is now a necessity. Many schools balk at more subjective evaluations because of legitimate equity concerns and legal risks, but rewarding "A" (count) while hoping for "B" (quality) will only become a bigger problem as the costs of "A" continue to fall.

Leadership at the department, school, and university levels must also create the conditions for these shifts to take hold through the recruitment and hiring process. If the hiring process continues to elevate publication counts and journal rankings as the currency of academic success, committees will optimize accordingly, regardless of what any editorial or task force recommends. Increased expectations that doctoral students have top publications when first entering the job market set the wrong norm at the beginning of careers, particularly when the quality of solo-authored job market papers (and the signal they send) suffers. As we teach our students, during periods of systemic change, over-anchoring on hard metrics risks diverting scholars from ambitious, frontier-expanding work to more derivative papers.

Given our evidence that AI-generated writing is making papers worse, authors need to consider the ethics of knowingly submitting *more* inferior papers rather than dedicating time to producing a single *better* manuscript. There is a real human cost to evaluating the growing volume of low-quality papers.

5.5. Conclusion

Our analysis of *Organization Science* shows that, at least in our corner of the social and behavioral sciences, we are heading toward a "more" rather than a "better" equilibrium as we integrate AI into our field. However, as researchers whose own work has been reshaped by AI over the past year, we also see the potential of a "better" equilibrium in which new, more difficult questions are asked, new methods are applied in imaginative ways, or both. We are encouraged by this potential, but concerned about the field's current trajectory. Whether our field evolves toward "more" or "better" is an institutional question, and will require us to adapt the incentives and culture that have allowed academics to thrive to this point.

Acknowledgments

Author names are in alphabetical order. All authors contributed equally to the letter. Gartenberg and Murray are Senior Editors, Hasan is a Deputy Editor, and Pierce is Editor-in-Chief of Organization Science. Hasan served as Chair of the journal's AI Task Force. The authors gratefully acknowledge support from the Wharton School at the University of Pennsylvania and the Fuqua School of Business at Duke University that enabled this analysis. The authors also thank Pangram for providing research credits and technical support during this effort. The article benefited from the comments of Seth Carnahan, Scott Dyreng, Frank Levy, Andrew Nelson, Oren Reshef, Christopher Tang,

Shirley Tang, and Matthew Walls. This project was submitted to the Duke University Institutional Review Board (IRB) for review, which ruled that it does not constitute human subjects research, as it is intended for institutional improvement. INFORMS legal counsel verified that the authors fully complied with the INFORMS Privacy Policy. AI was used in the following ways: Claude Code helped write Python scripts to merge data, calculate readability metrics, send batches to Pangram, and parse manuscripts into sections. Claude Code also helped write the Stata code for the analysis, particularly the figures. All code was reviewed by the authors before finalizing the analyses in the paper. Finally, the authors used Claude.AI to outline the discussion section and smooth out awkward phrasing. Grammarly was used for limited copyediting. All errors belong to the authors.

Endnotes

[1] We report only highly aggregated results in accordance with INFORMS data confidentiality and privacy policies.

[2] The ScholarOne database in which these are stored does not allow for automated downloads of the nearly 20,000 PDF documents necessary to evaluate all full-texts and reviews. In addition, the token usage for such a task would require an additional budget of $10,000.

[3] Complete descriptive statistics for the analysis sample are in Online Appendix A1.

[4] See https://www.pangram.com/research/model-card/pangram-3-1.

[5] Some prior research (Liang et al. 2025) suggests that AI detection algorithms mis-classify text written by non-native English speakers at higher rates. We find little evidence of this in a supplementary test in Online Appendix A2. Pre-ChatGPT texts are almost uniformly classified as "low AI" regardless of author origin, barring some small amount of model noise.

[6] Figure A2 in the Online Appendix shows a histogram of topic distributions for reviews flagged as low- versus high-AI usage.

[7] Though one could interpret less hedging and more active writing as overconfidence in the context of academic writing.

[8] This suggests that while AI use predicts desk rejection cross-sectionally, this effect disappears with author fixed effects. These two estimates are confounded. Authors who rely on AI may do so to address problems with inherently lower-quality papers, making it impossible to distinguish the effect of AI use from that of an author's manuscript quality, or from potential bias in decisions based on author identity. Nevertheless, the question of whether AI's effect on rejection rates reflects the technology itself or the types of authors who adopt it yields similar consequences for the journal: *Organization Science* is receiving too many AI submissions.

[9] In further robustness tests, we found that the results remained similar if we used an indicator for whether a manuscript received at least one 70%+ AI review. Thus, even the most AI-reliant reviews do not appear to be meaningfully shifting editorial decisions.

[10] While AI reduced the cost of writing manuscripts for the journal, the total cost of evaluating them has increased. This redistribution of complexity problem, as argued by Hasan et al. (2025), is what many organizations implementing or contending with AI will have to manage as AI shifts effort across tasks inside organizations.

[11] Interestingly, the issue of whether we are actually measuring "writing quality" came up in nearly all of the reviews we received in preparing this manuscript. Readers were surprised by the marked decline in writing quality scores after the launch of ChatGPT and proposed a range of alternative explanations. For instance: The measures we use are poor proxies for writing quality; poor writing reflects on the authors, not the AI; AI improves writing for non-Native speakers but not others; and so on. What is clear is that all traditional writing quality metrics show a significant decline with the use of AI. This appears to bump up against a very strong prior in our field that AI improves writing quality. This is a mismatch between priors and data worth exploring further, but it may be outside the scope of this article.

[12] Much of this high score is driven by the AI-generated writeup of definitions of the writing quality scores, which described the calculations made by the Python code.

[13] Their experimental evidence shows that AI use inflates authors' confidence regardless of actual quality. This is particularly concerning in an academic context, where the iterative struggle of writing is itself a form of thinking. When authors cede that process to AI, they may not realize that the resulting text, while fluent, lacks coherence and depth. This dynamic may be what we are capturing in manuscripts and reviews with Pangram scores above 30%.

References

Anthony C (2021) When knowledge work and analytical technologies collide: The practices and consequences of black boxing algorithmic technologies. *Admin. Sci. Quart.* 66(4):1173–1212.

Bechky BA, Davis GF (2025) Resisting the algorithmic management of science: Craft and community after generative AI. *Admin. Sci. Quart.* 70(1):1–22.

Dell'Acqua F, Ayoubi C, Lifshitz H, Sadun R, Mollick E, Mollick L, Han Y, et al. (2026) The cybernetic teammate: A field experiment on generative AI and teamwork. *Organ. Sci.* Forthcoming.

Dell'Acqua F, McFowland E III, Mollick E, Lifshitz H, Kellogg KC, Rajendran S, Krayer L, Candelon F, Lakhani KR (2026) Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organ. Sci.* 37(2):403–423.

Hasan S, Oettl A, Samila S (2025) From model design to organizational design: Complexity redistribution and trade-offs in generative AI. Preprint, submitted June 10, https://arxiv.org/abs/2506.22440.

Jabarian B, Imas A (2025) Artificial writing and automated detection. NBER Working Paper No. 34223, National Bureau of Economic Research, Cambridge, MA.

Jones N (2026) Leading preprint server clamps down on AI slop. *Science* 391(6784):432–433.

Kusumegi K, Yang X, Ginsparg P, de Vaan M, Stuart T, Yin Y (2025) Scientific production in the era of large language models. *Science* 390(6779):1240–1243.

Leonardi PM, Leavell V (2026) Knowing enough to be dangerous: The problem of "artificial certainty" for expert authority when using AI for decision making and planning. *Organ. Sci.* 37(2):516–543.

Liang W, Zhang Y, Wu Z, Lepp H, Ji W, Zhao X, Cao H, et al. (2025) Quantifying large language model usage in scientific papers. *Nature Human Behav.* 9:2599–2609.

Munger K, Bakker B, Berinsky A, Giger N, Guess A, Just N, Lawrence R, Tenenboim-Weinblatt K, van de Rijt A (2026) Peer review 2027: Scenarios for academic publishing in the age of AI. Preprint, submitted January 27, https://doi.org/10.31235/osf.io/594zj_v1.

Naddaf M (2025) Major AI conference flooded with peer reviews written fully by AI. *Nature* 648(8093):256–257.

Shaw SD, Nave G (2026) Thinking-fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender. Preprint, submitted January 11, https://doi.org/10.2139/ssrn.6097646.