Who gets to propose novel scientific ideas?*

Cristian Alamos†

November 21, 2026

Click here for the latest version

Abstract

Novel ideas drive scientific progress, yet their value is particularly difficult to judge in advance. The scientific community may therefore rely on authors' reputations to evaluate novel ideas, undervaluing those proposed by less established researchers. This paper studies whether novel ideas are underrated by peer review relative to their impact, especially when proposed by junior scientists. I extract referees' assessments of novelty from 123,000 reports on all submissions to a top artificial intelligence conference, measuring impact through citations and code reuse. Novel papers receive *higher* overall referee ratings, holding impact constant. However, this "novelty boost" is entirely absent for papers by junior authors. Using referees' self-reported confidence in their reports, I develop and estimate a Bayesian model of (editorial) decision-making under uncertainty, isolating editors' preferences for novelty from risk aversion and accurate statistical discrimination. Structural estimates imply that incorrect beliefs or seniority bias fully explain why the novel ideas of junior authors are relatively undervalued.

*I thank Lukas Althoff, Nicholas Bloom, Caleb Brobst, Sydnee Caldwell, Juliette Coly, Kathryn Cranney, Mi Liu, Adithya Mohan, Gideon Moore, León Tagle Le-Blanc, and participants at the SIEPR Predoctoral Research Salon for their insightful comments and feedback. The online appendix can be accessed here.

†Stanford Institute for Economic Policy Research. alamos@stanford.edu.

1 Introduction

Peer review shapes which scientific ideas are published and how scientific funding is allocated. Its judgments therefore influence the pace and direction of science—and with it, the supply of new ideas driving long-run economic growth. Yet, many scientists argue that peer review is biased against precisely the kind of ideas that matter most for progress: novel ones. Junior researchers, in particular, may lack the reputation that earns established researchers the benefit of the doubt when proposing pathbreaking ideas. These worries have been voiced by leading scientists, echoed in policy reports, and cited as justification for billions of dollars in alternative science funding.[1]

At the same time, a competing view holds that journals place too much emphasis on novelty, rewarding new eye-catching results with too little regard for whether they are well supported.[2] Moreover, trusting the ideas of more established authors may be rational if their track record provides a valuable signal of their credibility, especially for novel ideas, which may be more difficult to evaluate.

Despite the ubiquity of these concerns, there is little evidence on whether they are warranted. This is largely because these questions demand data that are difficult to obtain: referee evaluations of rejected work (which journals and funders rarely disclose) and reliable measures of how novel and difficult to evaluate papers are.

I study whether novel ideas are underrated by peer review relative to their impact, especially when proposed by junior researchers. I then examine the extent to which any effects can be explained by accurate statistical discrimination or by referees' and editors' attitudes towards risk.

Data. To answer these questions, I assemble a unique dataset of all submissions to the International Conference on Learning Representations (ICLR), one of the top three venues in artificial intelligence.[3] This conference makes the complete record of peer review publicly available for both accepted and rejected papers. The data cover 32,000 submissions and 123,000 referee reports from 2017 through 2025, including referees' ratings of each paper's quality on a 1–10 scale and the full text of their reports. I link these data to citations from papers and patents, to code-sharing data from GitHub, and to information on authors' job titles, publication records, and institutional affiliations.

These data allow me to measure four key variables necessary to answer my research questions: how difficult a paper is to evaluate, how novel it is, how impactful it is, and how established its authors are. First, to measure how difficult a paper is to evaluate, I rely on referees' self-reported

[1] See, e.g., Alberts et al. (2014), Braben et al. (2014), and Stephan et al. (2017) for scientists' views, National Academies (2007) and American Academy of Arts and Sciences (2008) for policy reports, and Sekar and Gallo (2023) and European Commission (2005) for funding.

[2] See, e.g., Nosek et al. (2012), Coffman et al. (2017), and Lipton and Steinhardt (2019) for this view, and Open Science Collaboration (2015) and Serra-Garcia and Gneezy (2021) for related evidence.

[3] ICLR has the second-highest h5-index among all artificial intelligence publications (Google Scholar, 2025). CSRankings, a widely used ranking of computer science departments, also counts ICLR among the three top venues in machine learning (Berger, 2026).

confidence in each evaluation on a 1–5 scale, which I reverse into a measure of the referee's uncertainty, from 1 ("absolutely certain") to 5 ("educated guess"). Throughout the paper, I use the term "uncertainty" to refer to the difficulty of evaluating a paper. I confirm that these ratings track the difficulty of evaluation: the share of the variance in citations explained by a referee's rating of overall quality is over twice as large for certain as for uncertain referees. Second, to measure novelty, I rely on the text of referee reports, using a large language model (LLM) to detect when a referee speaks positively or negatively about a paper's novelty. I show that these measurements agree with those of humans 94% of the time. In addition, my results are robust to using simple methods like counting the instances in which each referee uses the word "novel" in a positive light, and to using bibliometric proxies for novelty commonly used in the literature, such as how unusual the combinations of journals in its references are. Third, I measure impact via citations from papers and patents, as well as code-sharing data from GitHub. Fourth, I measure whether an author is established using their job titles, publication records, and institutional affiliations.

Descriptive Results. I begin with three descriptive findings. First, papers that referees describe as novel are accepted roughly twice as often as papers they describe as not novel. Second, novel papers receive 39% more citations than papers lacking novelty. Third, novel papers are substantially harder to evaluate: referees report significantly greater uncertainty (by 0.35 standard deviations) when evaluating novel papers.

Model. These findings motivate two questions. First, although referees and editors value novelty, do they value it enough given its relationship to long-run impact? Second, to the extent that novel ideas are undervalued or overvalued, is this explained by the greater difficulty of evaluating novel work, or by accurate statistical discrimination? I develop a Bayesian model of the editorial process to answer both.

Under this approach, novel papers are underrated if they achieve a higher average impact than other papers with the same rating. The intuition for this is as follows. Assume there are two papers whose overall ratings of quality are identical, but one is deemed novel and the other not novel. If evaluators weigh novelty according to an impact-maximizing benchmark, the two papers should go on to perform equally well on average. If instead the more novel paper has a higher impact on average, then evaluators are underrating novel papers, relative to their impact. This is because novel papers must achieve a higher average impact in order to receive the same overall ratings of quality.

In the model, each paper is characterized by its novelty and residual quality, neither of which evaluators observe directly. The referee receives noisy signals of these two components and uses them, along with their preference for novelty, to predict the quality of the paper in a Bayes-optimal manner. In addition, referees may have a taste for, or dislike of, papers that are difficult to evaluate, causing them to shift their impression of the overall quality of a paper. Thus, the model allows for two effects of difficulty of evaluation on how a paper is judged: it shrinks estimates of quality towards a prior mean, and it shifts perceptions of quality according to the referees' preference for

papers that are difficult to evaluate. Referees then report their shrunken estimates of novelty and overall quality to an editor. The editor then combines the referee reports with a private signal and decides whether to accept the paper, potentially weighting novelty differently than referees do. Once a decision is made, the paper's long-run impact again depends on its novelty and residual quality, as well as whether or not it is published in the conference.

An important property of this model is that the recovered parameters will not reflect accurate statistical discrimination. This is because, as long as referees optimally shrink the signals they report to editors, the recovered preferences for novelty will be the same as those reported under perfect information about a paper's quality (where there is no room for accurate statistical discrimination).[4] Importantly, this still leaves room for *inaccurate* statistical discrimination, where referees and editors have mistaken beliefs about the quality of novel papers.

A key challenge is that acceptance at a top venue mechanically increases a paper's citations—an effect I refer to as the "publication premium". I take three approaches to isolate the role of this premium. First, I measure impact using citations accrued before editorial decisions, where no publication premium can operate. Second, I estimate the publication premium directly through an event study comparing citation trajectories of borderline accepted and rejected papers. Third, I derive bounds from the model. Intuitively, if two papers receive the same referee scores but only one is accepted, the accepted paper will receive more citations partly because of the publication premium and partly because the editor had reason to prefer it. Thus, the total difference in citations between the two papers serves as an upper bound on the publication premium. A natural lower bound is zero, assuming that acceptance does not decrease a paper's impact. In practice, these bounds are tight.

Results. I find that referees and editors overvalue novel ideas relative to their eventual impact. Specifically, if two papers are given the same overall rating by a referee and one is described as novel and the other as lacking novelty, then the novel paper will on average receive 17% fewer citations from other papers, 16% fewer citations from patents, and its code will be reused by other researchers 19% less often. I refer to this effect as a "novelty boost". Editors correct for this, but only partially: their novelty boost is only a tenth to a quarter lower. This effect is also present when novelty is measured using features of the papers themselves: how far a paper's title and abstract are from the most similar earlier paper (Kelly et al., 2021; Arts et al., 2025); how unusual the pairs of journals in its reference list are (Uzzi et al., 2013); and how widely its references range across distant subfields (Leahey et al., 2017).

The novelty boost, however, is entirely absent from papers that are written by students or postdocs. The consequences of this are significant: if editors provided junior authors with the same novelty boost as for senior authors, novel papers by junior researchers would be accepted 29% more often. Given that acceptance at this conference nearly doubles the probability that a junior researcher moves to a top technology firm within three years (Miserocchi et al., 2026), this

[4] Appendix D formally tests and fails to reject the hypothesis that referees are acting in a Bayes-optimal manner.

can profoundly shape the career trajectories of junior researchers. The novelty boost is also twice as large when an author's H-index is in the top 10% of that year's submissions, and 77% larger when an author is at a top-20 institution.

Mechanisms. Why might referees overrate novel papers, relative to their eventual impact? As mentioned, the model implies that these results cannot be explained by accurate statistical discrimination. This means that referees either have incorrect beliefs about novel papers' eventual impact, or are not optimizing for average impact. I test and reject four natural explanations for why referees may be optimizing for a notion of quality not well-captured by average impact.

First, referees may be looking for "big hits"—papers with the potential to have a very large impact on their field—rather than papers that maximize expected impact. To test this, I ask whether novel papers are as likely as other papers with the same referee ratings to become one of the most cited submissions of their year. They are not: novel papers are 15% less likely to be in the top 20% of citations, and fully 47% less likely to be in the top 1%. Second, novel and difficult-to-evaluate papers may be "ahead of their time" and thereby attain a level of impact that reflects their true quality only in the long run. I evaluate this by following papers for five years and measuring the citations they receive in each year after the conference. The novelty boost is about as large in the fifth year as in the first, implying that this effect does not play a substantial role. Third, novel papers may be more influential to the papers that cite them. I test this by using Semantic Scholar's "influential" citations, which capture whether the citing paper builds substantially on the cited work. The novelty boost remains nearly unchanged when considering only influential citations. Fourth, novel ideas may be overrated because they are difficult to evaluate. In practice, controlling for how difficult a paper is to evaluate decreases the novelty boost by at most 13%.

Taken together, these results rule out the most natural rational explanations for why referees and editors may overrate novel ideas relative to their impact. They therefore suggest that referees and editors may either have incorrect beliefs about the quality of novel papers, or value novelty intrinsically above and beyond what can be justified by many notions of downstream impact.

Related Literature. Prior work has documented that novel ideas are cited more often (Uzzi et al., 2013; Wang et al., 2017) and get lower scores from grant evaluators in some settings (Boudreau et al., 2016; Ayoubi et al., 2021) but higher acceptance or renewal rates in others (Teplitskiy et al., 2022; Azoulay and Greenblatt, 2025). These studies rely on bibliometric proxies for novelty, such as unusual combinations of references or keywords. A smaller literature observes how experts judge novelty directly: panels selecting R&D projects in a firm (Criscuolo et al., 2017), evaluators of grant proposals in a field experiment (Lane et al., 2022), and NIH reviewers scoring the innovativeness of grant applications (Eblen et al., 2016). These studies, however, do not relate the experts' judgments to the later impact of the work. I make two primary contributions. First, I go beyond documenting associations and develop a model of the review process to determine whether novel ideas are underrated *relative to their impact*. Second, I explicitly consider the role of

uncertainty and of accurate statistical discrimination in driving this effect.[5]

A second literature studies whether peer review holds authors to different standards depending on their gender, connections, or prominence. In economics, prior work shows that an editor's colleagues publish more in the editor's journal, though their papers also tend to be cited more (Laband and Piette, 1994; Brogaard et al., 2014; Colussi, 2018), and that papers by women face a higher bar for publication (Card et al., 2020; Hengel, 2022). Experiments randomizing author names find that established authors are held to a lower standard (Okike et al., 2016; Tomkins et al., 2017; Huber et al., 2022), yet papers by prolific authors receive more citations than others with the same referee recommendations (Card and DellaVigna, 2020). I extend this work in two directions. First, because I observe referees' confidence in their evaluations, I can examine whether their attitudes towards the novel ideas of junior authors are driven by the fact that these ideas are more difficult to evaluate. Second, I show that these biases are exacerbated for papers that are novel and difficult to judge.

2 The ICLR Peer Review Process

The International Conference on Learning Representations (ICLR) is an annual artificial intelligence (AI) conference and one of the premier publication venues in computer science, a field where conferences serve as the primary outlets for research.[6] In 2025, it received over 11,000 submissions and had an acceptance rate of 32%, which is comparable to other top conferences in AI.[7]

The peer review process at ICLR involves two main roles: area chairs and reviewers. To follow the standard terminology in economics, I will refer to area chairs as editors and reviewers as referees, as their roles are analogous. Typically, 2–4 referees are assigned to each paper.

Referees are instructed to rate a paper's quality based on whether it "will bring sufficient value to the [AI] community and contribute new knowledge" (ICLR, 2024). In particular, they must write a referee report and rate the paper's overall quality on a scale from 1 to 10. Additionally, referees must report their confidence in their evaluation on a scale from 1 to 5.

Editors make acceptance recommendations based on referee reports and their own assessments. In practice, papers with a mean rating below 5 are almost always rejected, while those with a mean rating above 7 are accepted on nearly all occasions. In the "borderline" range of 5 through 7, where roughly 45% of papers lie, editors have the greatest scope to exercise discretion.

While the ICLR peer review process has been nominally double-blind since 2018, the identity

[5] I also show that common bibliometric proxies correlate only weakly with the referees' assessments of novelty in my data, contributing to a small literature on validating bibliometric proxies using the judgments of experts (Bornmann et al., 2019; Fontana et al., 2020).

[6] ICLR has the second-highest h5-index for artificial intelligence conferences according to Google Scholar (Google Scholar, 2025). The h5-index is defined as the maximum value of h such that the conference has published h papers that have each been cited at least h times during the last 5 years.

[7] In 2025, NeurIPS and CVPR had acceptance rates of 24.5% and 22.1% respectively.

of authors is frequently known by the referees and editors, who may have seen unanonymized pre-prints of the paper. If referees have not previously seen the paper, other revealing information may be included in the manuscript, acknowledgments, or linked code repositories. In small fields, the nature of the project itself may be sufficient information for de-anonymization (Bharadhwaj et al., 2020). Even when referees cannot discern the authors' identities from the manuscript, they may still search online for pre-prints. At another top conference in machine learning, 36% of referees who responded to a survey reported searching online for the papers they were reviewing (Rastogi et al., 2022).[8]

3 Data, Validation, and Descriptive Statistics

To study whether novel ideas are underrated by peer review, especially when proposed by junior authors, I measure four key variables: how novel a paper is, how difficult it is to evaluate, how much impact it has, and how prominent its authors are. I assemble these data by linking the complete record of peer review at ICLR from OpenReview to paper citations from Google Scholar, patent citations from Reliance on Science (Marx and Fuegi, 2020, 2022), code-sharing data from GitHub, authors' job titles and affiliations from OpenReview, and their publication records from Semantic Scholar.

3.1 Measuring Novelty

To measure novelty, I provide an LLM with the full text of a referee report and prompt it to ascertain whether the referee's assessment of the overall novelty of the paper is negative, neutral, or positive. If a report does not discuss novelty, it is labeled as such. I instruct the LLM to consider only explicit mentions of a paper's novelty, and to distinguish this from the general sentiment of the report towards other aspects of the paper.[9] In the estimation sample, 69% of reports assess novelty, and 97% of papers have at least one report that does.

This approach adds to a rapidly growing practice in economics of prompting an LLM to read a document and score a specific attribute of interest.[10] Asirvatham et al. (2026) find that, across more than 1,000 human-annotated benchmark examples spanning many domains, LLM-based ratings are "generally indistinguishable" from human evaluators.

A key advantage of this method is that the model observes the entire report, and therefore interprets each remark about novelty in the context of the entire review, much as a human reader would. A drawback is that this method lacks transparency, as the assessment cannot be traced

[8] The conference was ICML. The precise phrasing of the question was "During the review process, did you search for any of your assigned papers on the Internet?". 16% of referees responded to the survey.

[9] The model used is Claude Sonnet 5.5. The full text of the prompt is reproduced in Online Appendix OA.1.

[10] For instance, LLMs have been used to extract planned investment from earnings calls (Jha et al., 2024), the stance of Federal Reserve announcements (Hansen and Kazinnik, 2023), and the tone of news headlines (Lopez-Lira and Tang, 2026).

back to a single part of the report. Accordingly, I show in Online Appendix OA.2 that my results are robust to using an alternative measure based on identifying explicit mentions of the word "novel" and seeing if it is used in a positive light.

I validate this measure in four primary ways. First, I construct a human-labeled random sample of 200 referee reports that identifies whether the referee's assessment of the paper's novelty is negative, neutral, or positive. The human-labeled sample accords with the LLM's assessment 94% of the time. Second, I apply the classification measure to the text of 22,369 referee reports from the 2025 NeurIPS conference, which asked all its referees to rate the "originality" of each paper on a scale of 1 ("poor") to 4 ("excellent"). Among reports rating a paper's originality as "excellent", 95.1% are classified as positive and only 3.1% as negative. Among reports rating originality as "poor", 88.9% are classified as negative and 7.2% as positive. Third, I address the natural concern that this measure simply captures the overall sentiment of the review. The measure tracks referees' originality scores more closely than their overall assessment of the paper's quality: its correlation with originality is 0.54, against 0.27 for ratings of overall quality. Conditional on the referee's originality score, its partial correlation with the quality score is 0.06. Fourth, running the same prompt a second time on 600 randomly chosen reports yields scores that correlate at 0.94 with the first run.

3.2 Measuring Difficulty of Evaluation

I measure how difficult a paper is to evaluate through referees' self-reported confidence in their evaluations on a scale of 1–5. I reverse this scale and refer to it as the referee's *uncertainty*. Throughout the paper, a paper is said to be more difficult to evaluate when referees are more uncertain about it, and I use the two terms interchangeably. Referees generally describe themselves as certain: 13% of reports have an uncertainty of 1 ("absolutely certain") and 48% an uncertainty of 2 ("confident but not absolutely certain"). A further 32% have an uncertainty of 3 ("fairly confident"), with only 7% having an uncertainty of 4 or 5, indicating they likely "did not understand central parts of the paper". Figure B.1 shows that the ratings of certain referees are better predictors of eventual citations, of the ratings of other referees on the same paper, and of the editor's acceptance decision.

3.3 Measuring Impact

Citation Data from Google Scholar. My primary measure of impact is the number of citations a paper receives on Google Scholar. I link 97.3% of submissions from 2017 to 2025 to a Google Scholar record by searching for each paper's title and authors.[11] Submissions without a matching record are dropped from the sample. Among linked papers, the median has 9 citations and 16% have 0 citations.

[11] Citations were collected as of August 2026. The linking procedure is described in Online Appendix OA.1.

Code-Sharing Data from GitHub. As an alternative measure of impact, I track how frequently a paper's code has been adopted and modified by others. Specifically, I collect the number of times a paper's code repository has been forked on GitHub, a software sharing platform. Forks are copies of the code repository created by other users to modify or extend the original work.[12] This measure has the advantage of capturing an arguably more direct notion of impact than citations. Citations might not indicate that a paper was directly built upon, as researchers might cite to acknowledge results they merely mention, or to preempt referees who might object that a related paper was ignored. In contrast, forking a repository means the researcher took the authors' code in order to run, modify, or extend it. The drawback of this measure is that it is only available for 36.6% of submissions. Among papers with a repository, the median has 2 forks, and 36.9% of repositories have 0 forks.

Citations from Patents. As a further measure of impact, I use U.S. patent citations from Reliance on Science (Marx and Fuegi, 2020, 2022). This measure is only available for papers submitted between 2017 and 2019. I link 95.5% of these papers to Reliance on Science, and 41% of them are cited by at least one U.S. patent.

Influential Citations from Semantic Scholar. To capture whether later work builds substantially on a paper, I also use the citations that Semantic Scholar, a database indexing scientific papers and their citations, flags as "influential". Semantic Scholar classifies a citation as influential based on its context in the citing paper, such as how often and where the cited work is mentioned (Valenzuela et al., 2015).

3.4 Measuring Author Characteristics

To measure whether peer review is more open to the novel ideas of established authors, I use three measures of author prominence: H-index, institutional affiliation, and seniority. I measure each for the paper's most prominent author, and report results for the first and last authors in Appendix B.

H-Index. I define a paper as having an author who is prominent according to their publication record if the highest H-index among its authors is in the top 10% of that year's ICLR submissions. To measure H-indices, I link 96% of authors to a profile on Semantic Scholar and compute their H-index as of the year before they submitted to ICLR.[13] I show that my findings are robust to using alternative thresholds for H-indices in Table B.8.

Institutional Affiliation. I define an author as prominent according to their affiliation if they work at one of the twenty institutions with the most papers accepted at ICLR between 2017 and 2025.[14] To measure affiliations, I use OpenReview. All authors are required to create an account

[12] The matching procedure is described in Online Appendix OA.1.

[13] These metrics were collected in September 2026. The matching procedure and the construction of the as-of-submission measures are described in Online Appendix OA.1.

[14] These are Google (including Google DeepMind), Tsinghua University, Carnegie Mellon University, MIT, Microsoft,

with OpenReview upon submitting a paper, and are provided with the option of listing their job titles and affiliation history with dates, which gives me their job title and affiliation at the time of submission for 92% of authors.[15] I show that my results are robust to using university rankings from Times Higher Education and QS in Table B.8.

Seniority. I define an author as senior if they are neither a student nor a postdoc at the time of submission, based on the job titles they list on OpenReview. Roughly 36% of authors listed on papers are students, 5% are postdocs, 30% are faculty, and 29% are industry researchers.[16]

3.5 Measuring Referee and Editorial Perceptions of Quality

Ratings of Overall Quality. To measure how referees evaluate the quality of a paper, I use their ratings of overall quality on a scale from 1 to 10. A 3 is considered a "clear rejection", while scores of 5, 6, and 8 correspond respectively to papers that are "marginally below", "marginally above", and "clearly above" the acceptance threshold.[17]

Editorial Decisions. To measure how editors rate papers, I use their acceptance decisions. In the estimation sample, 33% of papers are accepted. Papers that authors withdraw after receiving their reports are treated as rejected. Desk-rejected papers receive no referee reports and are therefore not in the sample.

3.6 Additional Data

Subfields. I assign each paper to one of 21 subfields and include subfield fixed effects in all analyses. The taxonomy of subfields is derived directly from ICLR. Because submissions prior to 2025 were not assigned by authors to one of these 21 subfields, I use a large language model (Claude Haiku 4.5) to assign them based on their titles and abstracts. Online Appendix OA.1 lists the categories.

arXiv Pre-Prints. I link 58.5% of submissions to their arXiv pre-prints. This lets me determine how many citations papers received before the editorial decision. Online Appendix OA.1 describes how these citations are counted.

Meta, Stanford University, the University of California, Berkeley, Peking University, ETH Zurich, the University of Illinois Urbana-Champaign, Shanghai Jiao Tong University, Amazon, the Chinese University of Hong Kong, the University of Montreal, the University of Texas at Austin, the Hong Kong University of Science and Technology, Zhejiang University, the University of Oxford, and KAIST.

[15] Online Appendix OA.1 describes how author slots are linked to profiles and how the position at the time of submission is determined.

[16] These figures count an author once for every paper they appear on.

[17] The set of ratings referees may choose from varies across years: all ten values are available in 2017–2019 and 2021, only 1, 3, 6 and 8 in 2020, and only 1, 3, 5, 6, 8 and 10 from 2022 onward.

TABLE 1: Summary Statistics by Author Prominence

Sample:	All	Senior author		Top-20 institution		H-index top 10%	
		Yes	No	Yes	No	Yes	No
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Citations							
Asinh citations	2.87	2.89	2.37	3.27	2.51	3.51	2.82
	(1.96)	(1.96)	(2.04)	(2.00)	(1.85)	(2.02)	(1.94)
Top 5% most cited	0.05	0.05	0.02	0.08	0.03	0.10	0.04
Asinh influential citations	1.33	1.34	1.09	1.59	1.07	1.72	1.28
	(1.49)	(1.48)	(1.46)	(1.58)	(1.33)	(1.66)	(1.45)
Editorial decision							
Accepted	0.33	0.33	0.22	0.38	0.28	0.42	0.32
Referee reports							
Mean rating	5.22	5.24	4.74	5.43	5.05	5.52	5.21
	(1.26)	(1.24)	(1.45)	(1.21)	(1.26)	(1.19)	(1.25)
Referees per paper	3.85	3.85	3.70	3.84	3.86	3.86	3.84
Novelty							
Positive	0.34	0.34	0.29	0.35	0.33	0.36	0.34
Neutral	0.04	0.04	0.03	0.04	0.03	0.03	0.04
Negative	0.32	0.32	0.38	0.30	0.33	0.30	0.32
Not assessed	0.31	0.31	0.30	0.31	0.30	0.31	0.31
Difficulty of evaluation							
Very low	0.13	0.13	0.15	0.13	0.14	0.13	0.13
Low	0.48	0.48	0.46	0.49	0.48	0.50	0.48
Medium	0.31	0.31	0.31	0.32	0.31	0.31	0.32
High	0.07	0.07	0.08	0.07	0.07	0.06	0.07
Author prominence							
Senior author	0.97	1.00	0.00	0.98	0.96	0.98	0.97
Top-20 institution	0.47	0.48	0.31	1.00	0.00	0.72	0.45
H-index top 10%	0.10	0.10	0.06	0.15	0.06	1.00	0.00
Papers	32,161	30,793	1,073	15,148	16,748	3,252	28,496

Notes: This table presents the mean characteristics of all submitted papers, with standard deviations in parentheses. The sample is papers submitted to ICLR from 2017 to 2025, excluding 2020, and linked to Google Scholar, with each paper weighted equally. Columns 2–7 split papers by whether at least one author is senior (neither a student nor a postdoc), is at a top-20 institution, or has an H-index in the top 10% of that year's submissions; papers with an unknown characteristic are excluded from that pair of columns. Top 5% most cited is relative to submissions from the same year. Influential citations are from Semantic Scholar. Difficulty of evaluation is the referee's uncertainty: very low (1, "absolutely certain"), low (2, "confident but not absolutely certain"), medium (3, "fairly confident") or high (4 or 5, "educated guess").

3.7 Descriptive Statistics

Referees split almost evenly in how they discuss novelty: about a third of reports assess the paper's novelty positively, a third negatively, and a third do not assess it (Table 1). Most referees are certain about their evaluations, and only 7% report making an "educated guess".

Papers by prominent authors stand out on every measure of impact and evaluation. Papers

with an author at a top-20 institution, or with an author whose H-index is in the top 10%, receive more citations, are about twice as likely to be among the 5% most cited papers of their year, receive higher ratings, and are accepted about ten percentage points more often.

4 Model

To guide my analysis, I develop a stylized model of the editorial process. The model has four key properties that make it useful for my empirical analysis. First, it allows me to formally define what it means for an attribute like novelty or uncertainty to be "underrated" or "overrated" relative to a paper's impact, and to derive bounds on the size of that effect. Second, it shows how to recover editors' preferences for novelty and uncertainty, even though I do not observe their overall ratings of quality directly, as I do for referees. Third, it provides a framework for thinking about the role of uncertainty in evaluations. Uncertainty affects evaluations in two ways: it makes referees shrink their estimates towards a prior mean, and it may shift their assessment of quality directly, according to their taste for papers that are difficult to evaluate. Fourth, it recovers the parameters that referees and editors would have under perfect information, so by construction the estimated boosts do not capture accurate statistical discrimination. Importantly, this still leaves room for *inaccurate* statistical discrimination, where referees and editors have mistaken beliefs about the quality of novel papers.

The model works as follows. As illustrated in Figure 1, in the first step, the referee observes noisy signals of an attribute such as novelty n_R and residual quality r_R. They are aware of how noisy these signals are, and report their uncertainty about their evaluations U accordingly. They use these signals, as well as their preference for novelty β_R and for papers that are difficult to evaluate δ_R, to form estimates of overall quality $\hat{q}_R$ and novelty $\hat{n}_R$, which they report together with their uncertainty. The editor then combines $\hat{q}_R$ and $\hat{n}_R$ with their own private signals (n_E, r_E), as well as their preference for novelty β_E and for papers that are difficult to evaluate δ_E, to form a final acceptance decision. They accept a paper if their estimate of overall quality $\hat{q}_E$ is greater than a threshold τ.[18] All results are proven in Appendix A.

4.1 Setup

Let $U \in \{0, 1\}$ indicate whether the referee is highly uncertain about their evaluation ($U = 1$) or not ($U = 0$).[19] The referee's signals, defined below, are noisier when uncertainty is high. Conditional on U, let $n \sim \mathcal{N}(\mu_n(U), \sigma_n^2(U))$ denote a paper's latent attribute, such as novelty, and $r \sim \mathcal{N}(\mu_r(U), \sigma_r^2(U))$ denote residual quality independent of n given U. If referees and editors

[18] This is in contrast to Card and DellaVigna (2020), who treat referee recommendations as exogenous, who describe a setting where referees do not form or report perceptions of a noisily-observed characteristic like novelty, and who do not observe or model referees' uncertainty about their evaluations.

[19] For simplicity, I treat uncertainty as binary here. Appendix A allows it to take more than two values, as in my empirical specifications.

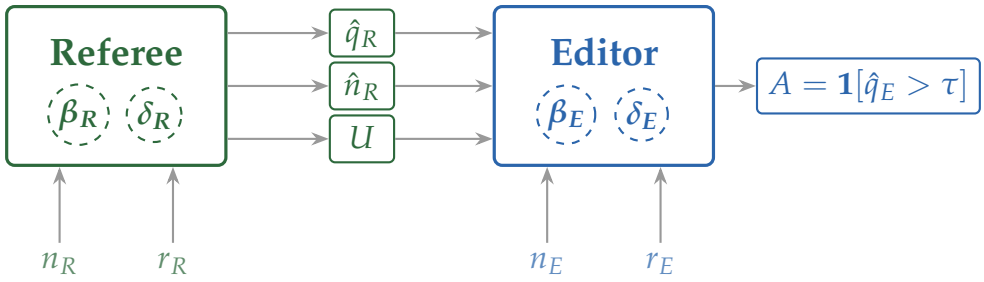


FIGURE 1: Model of the Editorial Decision Process

Notes: The referee observes noisy signals of novelty (n_R) and residual quality (r_R), and forms estimates of overall quality ($\hat{q}_R$) and novelty ($\hat{n}_R$), which they report together with their uncertainty (U), based on their preference weight β_R and their taste for papers that are difficult to evaluate δ_R. The editor then combines these referee reports with their own signals (n_E, r_E) and preference weight β_E and taste for papers that are difficult to evaluate δ_E to form a final acceptance decision. They accept a paper if their estimate of overall quality $\hat{q}_E$ is greater than a threshold τ.

observed n and r without noise, their assessments of quality would be

$$q_R = \beta_R\, n + r \;+\; \delta_R\, U, \qquad q_E = \beta_E\, n + r \;+\; \delta_E\, U, \tag{1}$$

where β_R and β_E respectively capture how strongly referees and editors value n, and δ_R and δ_E capture their taste for papers that are difficult to evaluate. In practice, they observe only noisy signals of both n and r:

$$n_R = n + \varepsilon_R, \quad r_R = r + \nu_R, \qquad n_E = n + \varepsilon_E, \quad r_E = r + \nu_E,$$

where ε_R, ν_R, ε_E, and ν_E are mean-zero Gaussian noises, mutually independent and independent of (n, r) given U. I denote the variances of the noise terms ε_R and ε_E by $\sigma_R^2(U)$ and σ_E^2, with $\sigma_R^2(1) > \sigma_R^2(0)$.

Lastly, let the impact of a paper be given by

$$q_I \;=\; \beta_I\, n \;+\; \psi\, A \;+\; r \;+\; \delta_I\, U \;+\; \epsilon, \tag{2}$$

where A is an indicator for whether a paper was accepted and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is independent of the latent variables and all signal noises. The parameters of interest are $\beta_I - \beta_R$ and $\beta_I - \beta_E$, which capture how much referees and editors undervalue or overvalue n relative to impact-maximizing behavior. I refer to these as the referees' and editors' novelty boosts, respectively. Analogously, $\delta_I - \delta_R$ and $\delta_I - \delta_E$ are referred to as uncertainty boosts. If these parameters are negative, referees and editors weight novelty and uncertainty more than they would under an impact-maximizing benchmark, meaning that they overrate these attributes relative to their eventual impact.

4.2 Solving the Model: Referee and Editorial Evaluations

Given their noisy signal n_R and their uncertainty U, the referee forms a shrunken estimate of n

$$\hat{n}_R \;=\; \mu_n(U) \;+\; \lambda_R(U)\,\big[n_R - \mu_n(U)\big], \qquad \text{where} \quad \lambda_R(U) \;=\; \frac{\sigma_n^2(U)}{\sigma_n^2(U)+\sigma_R^2(U)}.$$

Let $\hat{r}_R$ denote the referee's estimate of residual quality based on r_R, defined analogously. The referee's quality assessment is then

$$\hat{q}_R \;=\; \beta_R\,\hat{n}_R \;+\; \hat{r}_R \;+\; \delta_R\,U. \tag{3}$$

The editor then combines the referee's information with their own signals. The editor's assessment can be written as

$$\hat{q}_E \;=\; \beta_E(\theta(U)\,n_E + (1-\theta(U))\,\hat{n}_R) \;+\; \alpha(U)\,r_E \;+\; (1-\alpha(U))\,\hat{r}_R \;+\; \delta_E\,U, \tag{4}$$

where $\theta \in (0,1)$ governs how the editor weights their own signal of n, denoted n_E, against the referee's $\hat{n}_R$, and $\alpha \in (0,1)$ serves an analogous role for residual quality. The relative weight that the editor places on their own signal of novelty n_E and on the referee's signal n_R depends on how noisy each signal is:[20]

$$\frac{\theta(U)}{1-\theta(U)} \propto \frac{\sigma_R^2(U)}{\sigma_E^2}.$$

The editor accepts the paper if $\hat{q}_E > \tau$ for a finite threshold τ.

4.3 Model Estimation

I now turn to estimating the extent to which referees and editors deviate from impact-maximizing behavior in their evaluations of novel ideas and of ideas that are difficult to evaluate. This amounts to estimating $\beta_I - \beta_R$ and $\beta_I - \beta_E$, respectively, for novelty, and $\delta_I - \delta_R$ and $\delta_I - \delta_E$ for uncertainty. I describe the estimation process for the novelty boost; the parameters for uncertainty are estimated in an analogous manner from the coefficients on U.

4.3.1 The Referee's Novelty Boost

Re-arranging (2) and (3) to express impact in terms of observables yields:

$$q_I \;=\; (\beta_I - \beta_R)\,\hat{n}_R \;+\; (\delta_I - \delta_R)\,U \;+\; \hat{q}_R \;+\; \psi\,A \;+\; \nu. \tag{5}$$

[20] Here, $\theta(U) = \dfrac{\sigma_E^{-2}}{\sigma_n^{-2}(U)+\sigma_R^{-2}(U)+\sigma_E^{-2}}$, and the factor of proportionality is the referee's shrinkage factor $\lambda_R(U)$.

When a publication premium is present, estimating this equation through OLS will yield biased estimates of $(\beta_I - \beta_R)$, as acceptance decisions are endogenous if the editor has private information about paper quality (i.e., $\text{Cov}(A, \nu) > 0$).

Point Estimates. One approach to obtaining point estimates is to separately identify the causal effect of being published at ICLR on citations ψ and then estimate:

$$q_I - \psi A = b_0 + b_1 \hat{n}_R + b_2 \hat{q}_R + b_3 U + \nu_b. \tag{6}$$

As detailed in Appendix C, ψ can be identified through an event study around the decision deadline for papers posted on arXiv prior to editorial decisions. A limitation of this approach is that the estimation sample is limited to papers with average scores that put them on the margin of acceptance and that are posted on arXiv prior to the editorial decision. Given these limitations, it is useful to also estimate upper and lower bounds on novelty boosts.

Bounds. We can derive bounds on $\beta_I - \beta_R$ by regressing impact q_I on novelty $\hat{n}_R$, overall ratings $\hat{q}_R$, and uncertainty U, with and without an indicator A for whether or not a paper was accepted. Specifically, the coefficient on $\hat{n}_R$, c_1, in equation (7) will underestimate $\beta_I - \beta_R$ by $\beta_E - \beta_R$ times a positive constant proportional to $\text{Cov}(A, \nu)$. Conversely, if we remove the acceptance indicator from the regression as in equation (8), the coefficient on $\hat{n}_R$, d_1, will overestimate $\beta_I - \beta_R$ by a positive constant times $(\beta_E - \beta_R)$ (Proposition A.6).

$$q_I = c_0 + c_1 \hat{n}_R + c_2 \hat{q}_R + c_3 A + c_4 U + \nu_c, \tag{7}$$

$$q_I = d_0 + d_1 \hat{n}_R + d_2 \hat{q}_R + d_3 U + \nu_d. \tag{8}$$

Exactly which of c_1 and d_1 is the lower and upper bound depends on whether referees or editors value n more (i.e., whether $\beta_E > \beta_R$). In practice, these bounds are very tight.

The intuition for these bounds is as follows. Accepted papers will have higher impact conditional on referee ratings both because of the publication premium ψ, and also because editors have private information about quality beyond that contained in referee ratings. Thus, the coefficient c_3 in equation (7), which captures the effect of both the publication premium ψ and the editor's private signal of quality, provides an upper bound on ψ. A reasonable lower bound for ψ is 0, assuming that getting published at ICLR does not actively decrease a paper's expected citations. Thus, if $\psi = c_3$ then $\beta_I - \beta_R = c_1$, and if $\psi = 0$, then $\beta_I - \beta_R = d_1$. Because $\psi \in [0, c_3]$, $\beta_I - \beta_R$ is bounded by c_1 and d_1 (Proposition A.6).

4.3.2 The Editor's Novelty Boost

I now turn to the identification of the editor's novelty boost, $\beta_I - \beta_E$.

We can estimate a scalar multiple of $\beta_E - \beta_R$ by regressing the acceptance decision A on observable paper characteristics and referee ratings $\hat{q}_R$ using a probit model. To see this, notice that

we can re-arrange equation (4) to express $\hat{q}_E$ in terms of observables and $\hat{q}_R$ as

$$\hat{q}_E \;=\; (\beta_E - \beta_R)\,\hat{n}_R \;+\; (\delta_E - \delta_R)\,U \;+\; \hat{q}_R \;+\; \zeta,$$

where ζ is mean-zero Gaussian and independent of $(\hat{n}_R, \hat{q}_R, U)$ (Proposition A.3).[21] The acceptance rule can be written

$$A \;=\; \mathbf{1}[(\beta_E - \beta_R)\,\hat{n}_R \;+\; (\delta_E - \delta_R)\,U + \hat{q}_R + \zeta > \tau]\,,$$

which, recalling the normality assumption, implies the probit model

$$\Pr(A = 1 \mid \hat{n}_R, \hat{q}_R, U) \;=\; \Phi\left[\frac{(\beta_E - \beta_R)\,\hat{n}_R \;+\; (\delta_E - \delta_R)\,U + \hat{q}_R - \tau}{\sigma_\zeta}\right].$$

Therefore, estimating

$$\Pr(A = 1 \mid \hat{n}_R, \hat{q}_R, U) = \Phi\big[\gamma_0 + \gamma_n \hat{n}_R + \gamma_q \hat{q}_R \;+\; \gamma_U\, U\big] \tag{9}$$

will yield the coefficients

$$\gamma_n = \frac{\beta_E - \beta_R}{\sigma_\zeta}, \qquad \gamma_q = \frac{1}{\sigma_\zeta},$$

and hence,

$$\frac{\gamma_n}{\gamma_q} = \beta_E - \beta_R.$$

Therefore, dividing γ_n by γ_q identifies $\beta_E - \beta_R$. So far, I have treated ratings of quality as continuous, with a coefficient normalized to 1 in the citation regression. In practice, ratings are discrete, so I represent them nonparametrically, with an indicator for each rating, in both the citation regression and the probit. In the model, the probit coefficient on each rating equals its citation coefficient times σ_ζ^{-1}, so the ratio of the two recovers σ_ζ^{-1}. If this ratio is the same for every rating, this supports the key structural assumption that ratings of overall quality enter proportionally into the citation and probit regressions: plotting the probit coefficients against the citation coefficients for every rating, the points should lie on a straight line through the origin with slope σ_ζ^{-1} (Figure 2). By a similar logic to the previous subsection, estimating the citation regression with and without the acceptance indicator A yields upper and lower bounds on σ_ζ^{-1}, and subtracting the estimated publication premium $\hat{\psi}A$ from citations yields a point estimate. Finally, with σ_ζ^{-1} in hand, we can recover the editor's novelty boost $\beta_I - \beta_E$ directly by taking

$$(\beta_I - \beta_R) - \left(\frac{\beta_E - \beta_R}{\sigma_\zeta}\right)\sigma_\zeta = \beta_I - \beta_E. \tag{10}$$

In practice, I estimate the citation and acceptance equations jointly by maximum likelihood, im-

[21] Because the referee's noise varies with uncertainty, the variance of ζ, σ_ζ^2, could in theory also vary with uncertainty; empirically it does not (Appendix D).

posing that the probit coefficient on each rating equals its citation coefficient times σ_ζ^{-1}. Section 5.1 walks through both the step-by-step calculation and the joint estimates.

4.3.3 Bringing the Model to the Data

As detailed in Section 3, the dataset contains measures of novelty and referee uncertainty from referee reports, acceptance decisions A, impact q_I (e.g., asinh citations and GitHub forks), and referee ratings of overall quality $\hat{q}_R$. I therefore estimate the following equations for referee report j on paper i:

$$\text{asinh}(\text{Impact}_i) = \beta\,\text{Novelty}_{ij} + \alpha'\,\text{Rating}_{ij} + \gamma'\,\text{Uncertainty}_{ij} + \theta' X_{ij} + \varepsilon_{ij}, \tag{11}$$

$$\Pr(\text{Accepted}_i) = \Phi\big(b\,\text{Novelty}_{ij} + a'\,\text{Rating}_{ij} + g'\,\text{Uncertainty}_{ij} + t' X_{ij}\big). \tag{12}$$

Novelty_{ij} equals $+\frac{1}{2}$ if the referee's assessment of novelty is positive, 0 if neutral, and $-\frac{1}{2}$ if negative. Its coefficient therefore captures the difference in outcomes between a positive and a negative assessment. I add an indicator for reports that do not mention novelty. I treat ratings of overall quality $\hat{q}_R$ on a scale from 1 to 10 nonparametrically: Rating_{ij} contains indicators for ratings of 4, 5, 6, 7, and 8 or more. To capture how difficult a paper is to evaluate, Uncertainty_{ij} contains indicators for low (uncertainty = 2), medium (3) and high (4 or 5) uncertainty. X_{ij} contains fixed effects for year and subfield, and for the number of referees assigned to a paper interacted with the year. I estimate every equation at the level of the individual referee report, weight each report by the inverse of its paper's number of referees, and cluster standard errors at the paper level. In Section 5.3, I add indicators for author prominence to both equations and interact them with novelty and uncertainty.

5 Results

This section presents three sets of results. Section 5.1 establishes that novel ideas are overrated relative to their eventual impact. Section 5.2 examines the potential mechanisms driving this effect. Section 5.3 asks whether novel ideas by less established authors are held to a different standard.

5.1 Are novel ideas underrated relative to their eventual impact?

Referees. Referees tend to overrate papers they describe as novel, relative to their eventual impact. Specifically, estimating equation (11) shows that if two papers are given the same overall rating by a referee, and one is described as novel and the other as lacking novelty, then the novel paper will on average receive 15% to 17% *fewer* citations. In other words, novel papers face a lower bar for receiving the same overall rating of quality. This is not because novel papers are cited less overall: without controlling for referees' overall ratings, papers that a referee describes as novel

receive 39% *more* citations than papers described as lacking novelty (Table 2). Researchers also build on novel papers' code 19% less often, and novel papers receive 16% fewer citations from patents (Table B.1). The point estimate of the novelty boost, in column 3 of Table 2, is derived by subtracting the publication premium $\hat{\psi}$, estimated from the event study in Appendix C, from citations, as in equation (6). The bounds in columns 2 and 4 are derived by estimating equations (8) and (7), without and with an indicator for whether a paper was accepted. In practice, these bounds are remarkably tight: they range from -0.17 to -0.15, a difference of less than 15% of the point estimate.

Editors. Editors place less weight on novelty than referees do, but still overrate it relative to its eventual impact. As shown in column 5, which estimates equation (12), editors place negative weight on novelty when making their acceptance decisions, conditional on referees' overall evaluations of quality. This implies editors value novelty less than referees, and recovers the parameter $\frac{\beta_E - \beta_R}{\sigma_\zeta} = -0.062$ in the context of the model.

To establish whether editors still overrate novelty *relative to impact*, it is necessary to "convert" between probit and OLS coefficients. In the context of the model: now that we have estimates of $\beta_I - \beta_R$ and $\frac{\beta_E - \beta_R}{\sigma_\zeta}$, we need to estimate σ_ζ to recover $\beta_I - \beta_E = (\beta_I - \beta_R) - \left(\frac{\beta_E - \beta_R}{\sigma_\zeta}\right)\sigma_\zeta$. As shown in Section 4.3, σ_ζ^{-1} can be recovered as the ratio of the probit to the OLS coefficients on the indicators for referees' overall ratings.[22] Figure 2 plots the probit coefficients against the OLS coefficients. The slope of the resulting line provides the estimate for $\sigma_\zeta^{-1} = 1.48$. Thus, plugging in parameters from the point estimates of the novelty measure yields a point estimate of $\beta_I - \beta_E = -0.167 - (-0.062)/1.48 = -0.167 + 0.042 = -0.125$.

Structural Estimates. These calculations can be formalized by estimating the citation and acceptance equations jointly by maximum likelihood, imposing the structural assumption that referees' overall ratings enter proportionally in the citation and acceptance equations. This allows me to directly estimate $\beta_I - \beta_E$, with standard errors that account for the joint estimation of all three estimands: the referee's novelty boost, the editor's relative weight on novelty, and σ_ζ. Column 6 of Table 2 reports the resulting estimate: $\beta_I - \beta_E = -0.15$ (s.e. 0.01).[23] Treating the estimates with and without the acceptance indicator as bounds, the 95% confidence interval for this partially identified parameter, following Imbens and Manski (2004), is the bracket in column 6: $[-0.17, -0.12]$. Thus, if an editor faces two papers with the same referee ratings, and the referees deem one novel and the other not, the novel paper will receive 15% fewer citations on average: editors undo only a tenth to a quarter of the novelty boost that referees grant to novel papers.

Figure 2 visually illustrates that editors do not fully overturn the novelty boost of referees. The black line shows the coefficients on overall ratings of quality in the probit (y-axis) and OLS (x-axis) regressions. As shown in the model section, this line has a slope of $\sigma_\zeta^{-1} = 1.48$. If novelty

[22] The probit and OLS coefficients lie on a straight line through the origin across the rating distribution, supporting the model's assumption that ratings of overall quality enter proportionally into the citation and probit regressions.

[23] This is larger in magnitude than the back-of-the-envelope value of -0.13, because imposing proportional rating coefficients in both equations also changes the estimated novelty coefficients.

were weighted according to an impact-maximizing benchmark by both referees and editors, its coefficients in the probit and OLS equations would both be 0 and the blue dot would lie at the origin. If instead novelty is not weighted correctly by referees, but editors fully correct for this bias, then the blue dot would lie on the black line. Intuitively, if novelty is overrated by referees, its coefficient in the OLS equation will be negative. For editors to overturn this effect, they have to value novelty less than referees, meaning that the coefficient in the probit equation has to be negative. How negative does it have to be to exactly overturn the effect? As shown in the model, it has to be σ_ζ^{-1} times the OLS coefficient. Given that the slope of the line corresponds to σ_ζ^{-1}, this means it has to be on the black line. As shown in Figure 2, the blue dot lies well above the black line, and the slope of the blue line is below even the lower bound for the slope of the black line, implying that editors do not fully reverse the novelty boost.

Alternative Measures of Novelty. These results hold when using referees' own numerical ratings of originality on a scale from 1 ("poor") to 4 ("excellent") that the NeurIPS AI conference collected from every referee in 2025 (Table B.2). If a referee gives two papers the same overall rating, but deems one's originality "poor" (originality = 1), and the other's "excellent" (originality = 4), then the more novel paper will receive roughly 17% fewer citations. Because these scores are available only for accepted papers, and because the resulting sample is roughly nine times smaller than the ICLR sample, I use this measure only as a supplementary measure. The novelty boost also appears when novelty is measured from attributes of the papers themselves, as is common in the literature (Online Appendix Table OA.3). These measures capture how far a paper's title and abstract are from the most similar earlier paper (Kelly et al., 2021; Arts et al., 2025), how unusual the pairs of journals and conferences in its reference list are (Uzzi et al., 2013), and how widely its references range across subfields (Leahey et al., 2017). They also correlate positively with referees' assessments of novelty.

Uncertainty. Another interesting finding is that both referees and editors *overrate* ideas that are difficult to evaluate, relative to their eventual impact. If a referee gives two papers the same overall rating, but is "absolutely certain" in their evaluation of one (uncertainty = 1) and highly uncertain about the other (uncertainty of 4 or 5), then the paper they are uncertain about will on average receive 34% to 37% fewer citations (Table 2). The effect grows with the referee's uncertainty: it is 6% for low and 19% to 21% for medium uncertainty. I refer to this effect as an "uncertainty boost". Editors undo little of it: for high uncertainty, the joint estimate of $\delta_I - \delta_E$ is -0.30. This can also be seen visually in Figure 2, where the coefficients on uncertainty form a slope that is significantly less steep than even the lower bound for the slope of the black line. Papers reviewed with high uncertainty also receive fewer citations from patents, and their code is reused less often (Table B.1). In other words, referees and editors give papers they find difficult to evaluate the benefit of the doubt.

Robustness. These results are robust to dropping reports that do not mention novelty from the sample and to including papers that have no citation data and imputing their citations as 0 (Table B.1). They also hold when impact is measured with the citations a paper receives before the

editorial decision, where no publication premium can operate, and among papers with an average rating below 5, of which only 0.5% are accepted, so that there is almost no publication premium: in both cases, the novelty boost is 11% (Table B.1). They are also similar when measuring novelty using explicit mentions of the word "novel", which yields a significant novelty boost of 17% for editors (Online Appendix OA.2).

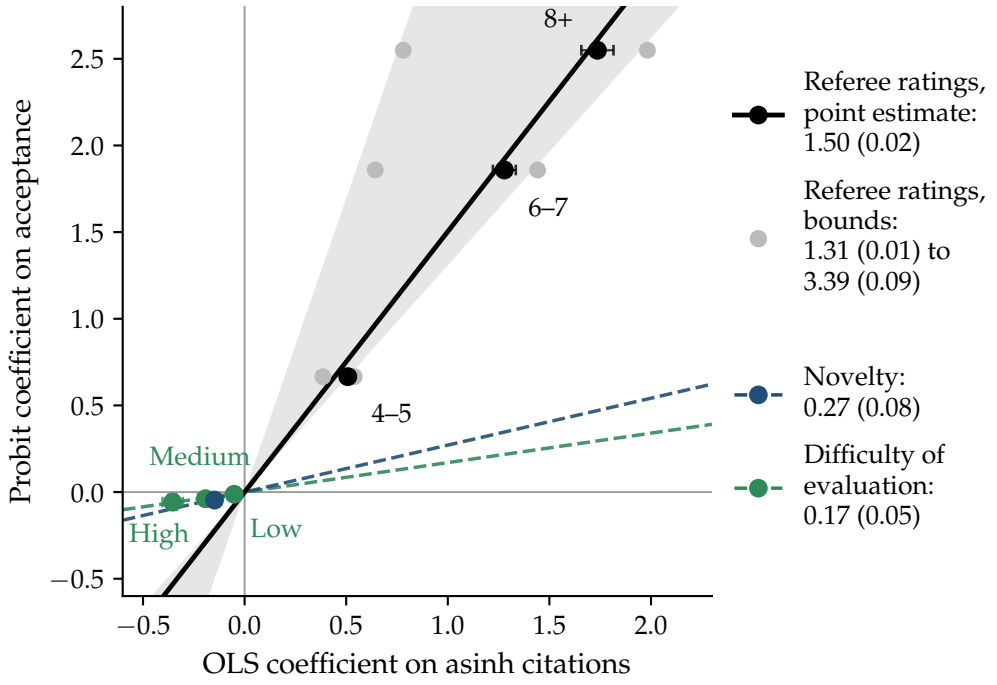


FIGURE 2: Coefficients for Referee Ratings, Difficulty of Evaluation, and Novelty

Notes: Each point plots a coefficient from the probit model for acceptance (Table 2) against the corresponding coefficient from the OLS model for the inverse hyperbolic sine of citations. Black dots represent point estimates of the coefficients on ratings of overall quality, from equation (6); gray dots represent the upper and lower bounds, from equations (8) and (7). The novelty and difficulty-of-evaluation coefficients are from the same regression as the black dots. Lines are joint maximum-likelihood fits through the origin. Bars represent 95% confidence intervals clustered by paper.

5.2 Mechanisms

Why might referees overrate papers that are novel and difficult to evaluate, relative to their eventual impact? As shown in the model section, this cannot be explained by accurate statistical discrimination. This means that referees either have incorrect beliefs about novel papers' eventual impact, or are not optimizing for average impact. I test and reject four natural explanations for why referees may be optimizing for a notion of quality not well-captured by average impact.

First, referees may be looking for "big hits"—papers with the potential to have a very large impact on their field—rather than papers that maximize expected impact. If novel and difficult-to-evaluate ideas have a higher variance of outcomes, then referees and editors could rationally prefer papers with these qualities. I find no evidence that these papers are disproportionately

TABLE 2: Models for Impact and Acceptance Decisions

		OLS: asinh citations ($\beta_I - \beta_R$)			Probit: accepted ($\frac{\beta_E - \beta_R}{\sigma_\zeta}$)	Joint MLE ($\beta_I - \beta_E$)
	(1) $\psi = 0$	(2) $\psi = 0$ (Bound)	(3) $\psi = 0.33$ (Point)	(4) $\psi = 1.58$ (Bound)	(5) – (Point)	(6) $\psi = 0.33$ (Point)
Novelty	0.39***	–0.17***	–0.17***	–0.15***	–0.06***	–0.15***
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	[–0.17, –0.12]
Difficulty of evaluation						
Low	0.09***	–0.06***	–0.06***	–0.06***	–0.02	–0.05**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	[–0.08, –0.01]
Medium	0.06**	–0.21***	–0.20***	–0.19***	–0.04**	–0.17***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	[–0.20, –0.13]
High	–0.07**	–0.37***	–0.36***	–0.34***	–0.06**	–0.30***
	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	[–0.34, –0.24]
Overall rating						
4		0.03	0.01	–0.06	0.31***	
		(0.05)	(0.04)	(0.04)	(0.04)	
5		0.62***	0.58***	0.45***	0.71***	
		(0.02)	(0.02)	(0.02)	(0.02)	
6		1.37***	1.22***	0.62***	1.82***	
		(0.02)	(0.03)	(0.02)	(0.02)	
7		2.16***	1.93***	1.05***	2.45***	
		(0.05)	(0.06)	(0.04)	(0.03)	
8+		1.99***	1.75***	0.81***	2.56***	
		(0.02)	(0.04)	(0.02)	(0.02)	
Accepted				1.58***		
				(0.02)		
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Subfield FE	Yes	Yes	Yes	Yes	Yes	Yes
Referee count × year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	123,666	123,666	123,666	123,666	123,666	123,666

Notes: The sample is reviews of ICLR 2017–2025 submissions, excluding 2020, linked to Google Scholar. Observations are weighted by the inverse of the number of reviews of the paper. The dependent variable is the inverse hyperbolic sine of Google Scholar citations in columns 1–4 and an indicator for acceptance in the probit model of column 5. Column 6 is the joint maximum-likelihood estimate described in Section 5.1. All models include an indicator for whether the report contains an assessment of the paper's novelty, as well as the paper's year of submission and subfield, and the number of referees interacted with the year of submission. Standard errors clustered by paper in parentheses; in column 6, brackets are 95% confidence intervals over the range of ψ (Imbens and Manski, 2004). $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

likely to lead to "big hits". To test this, I replace the outcome variable in the regressions of Table 2 with an indicator for whether a paper is among the top 1, 10, or 20 percent most cited submissions to ICLR that year. Conditional on referee ratings, novel papers are 15% less likely to be in the top quintile than non-novel papers, 22% less likely to be in the top decile, and fully 47% less likely to be in the top centile of citations among submissions to ICLR that year (Table B.3). The results for uncertainty are analogous: papers reviewed with high uncertainty are 38% less likely to be in the top quintile, 53% less likely to be in the top decile, and 69% less likely to be in the top centile than papers reviewed with very low uncertainty.

Second, novel and difficult-to-evaluate papers may be "ahead of their time" and thereby attain a level of impact that reflects their true quality only in the long run. To test this, I follow papers for five years and measure the citations they receive in each year after the conference. I restrict this analysis to papers submitted from 2017 to 2019, so that the same papers are observed in every year. The novelty boost and the uncertainty boost do not shrink as papers age: they are about as large in the fifth year as in the first (Figure B.2).

Third, novel and difficult-to-evaluate papers may be more influential to the papers that cite them. That is, these papers could receive fewer citations but be cited more substantively, for instance because later work builds on their methods rather than mentioning them in passing. To test this, I use Semantic Scholar's "influential" citations, which capture whether the citing paper builds substantially on the cited work. The novelty boost is only slightly smaller when considering only influential citations: novel papers receive 18% fewer influential citations than other papers with the same referee ratings, against 21% fewer citations overall in the same sample (Table B.4). Similarly, papers reviewed with high uncertainty receive 33% fewer influential citations than papers reviewed with very low uncertainty and the same referee ratings, against 44% fewer citations overall.

Lastly, novel ideas may be overrated because they are difficult to evaluate. At most 13% of the novelty boost is explained by the fact that novel ideas are more difficult to evaluate. Specifically, controlling for the referee's uncertainty about their evaluation reduces the coefficient on novelty by at most 13%.

Taken together, these results rule out the most natural rational explanations for why referees and editors may overrate novel ideas and ideas that are difficult to evaluate relative to their long-run impact. They therefore suggest that referees and editors may either have incorrect beliefs about the quality of these papers, or value novelty and uncertainty intrinsically above and beyond what can be justified by many notions of downstream impact.

5.3 Who gets to propose novel scientific ideas?

I now examine whether referees and editors are more lenient towards the novel ideas of more established authors. Specifically, having established that on average authors receive a "novelty

boost", I now ask whether this boost is higher for authors with high H-indices, authors at top institutions, or senior authors. Having walked through the step-by-step derivation of the structural estimates in Section 5.1, I now present structural estimates of editors' novelty boosts directly for brevity, and report the reduced-form estimates in Table B.6.

The novelty boost is completely absent for papers written entirely by students and postdocs, and is weaker for papers without highly published or top-affiliated authors (Table 3). Specifically, if two papers written entirely by students and postdocs receive the same referee ratings, and one is described as novel and the other as lacking novelty, then the novel paper will receive about as many citations on average. In contrast, if both papers have at least one senior author, then the novel paper will on average receive 14% fewer citations. In addition, papers with an author whose H-index is in the top 10% see a novelty boost that is roughly twice as large as other papers, while papers with an author at a top-20 institution see a novelty boost that is 77% larger. Referees show a similar pattern to editors (Panel B of Table B.7).

About half of the disproportionately large novelty boost of senior authors can be explained by the fact that senior authors are also more highly published, and more often at top institutions. Papers written entirely by students and postdocs are less likely to have an author whose H-index is in the top 10% of that year's submissions (6%, against 10% of other papers) and less likely to have an author at a top-20 institution (31%, against 48%). Table 3 therefore also estimates a model that includes all measures of author prominence at once: a top-10% H-index, a top-20 institution, and a senior author. In this regression, senior authors' larger novelty boost falls from 16 to 8 percentage points, with two thirds of this decline coming from the addition of a control for H-index. The larger novelty boosts of authors with a high H-index or at a top-20 institution, by contrast, are mostly not driven by the other measures of prominence: in the same regression, they fall by a fifth and by a third, respectively.

The results are similar when the characteristics are measured for the paper's *last* author, who in computer science is often the senior researcher supervising the work (Panel A of Table B.7). The novelty boost is absent when the last author is a student or postdoc, and it is roughly twice as large when the last author has an H-index in the top 10%. A last author at a top-20 institution raises the novelty boost by about 40%, although this difference is significant only at the ten percent level.

The publication record and seniority of the *first* author, on the other hand, do not seem to matter (Panel C of Table B.7). In computer science, the first author is often a junior researcher responsible for most of the work. Given that first authors are generally not expected to be senior, one natural interpretation of this result is that referees and editors do not weigh their seniority and publication record. Whether the first author is at a top-20 institution, however, does matter: these papers receive a novelty boost that is 42% larger.

Alternative Measures of Impact. The results are similar in magnitude when impact is measured with GitHub forks (Table B.9). However, forks are available for only a third of the papers, so most of the estimates are statistically indistinguishable from zero; only the larger novelty boost of

TABLE 3: Structural Estimates of Editors' Novelty Boost by Author Characteristics

	$(\beta_I - \beta_E)$			
	(1)	(2)	(3)	(4)
Novelty	–0.12***	–0.10***	0.02	–0.02
	[–0.15, –0.09]	[–0.13, –0.06]	[–0.13, 0.20]	[–0.18, 0.15]
× H-index top 10%	–0.12**			–0.10*
	[–0.23, –0.03]			[–0.20, –0.01]
× Top-20 institution		–0.07**		–0.05†
		[–0.14, –0.02]		[–0.12, 0.00]
× Senior			–0.16*	–0.08
			[–0.34, –0.01]	[–0.25, 0.07]
Difficulty of evaluation	–0.10***	–0.10***	–0.04	–0.07
	[–0.12, –0.08]	[–0.12, –0.08]	[–0.11, 0.05]	[–0.14, 0.02]
× H-index top 10%	–0.04†			–0.04†
	[–0.09, 0.00]			[–0.09, 0.01]
× Top-20 institution		–0.01		0.00
		[–0.04, 0.02]		[–0.03, 0.03]
× Senior			–0.07†	–0.04
			[–0.15, 0.00]	[–0.12, 0.04]
Rating dummies	Yes	Yes	Yes	Yes
Year and subfield FE	Yes	Yes	Yes	Yes
Referee count × year FE	Yes	Yes	Yes	Yes
N	123,666	123,666	123,666	123,666

Notes: See the notes to Table 2. Each column estimates the editor's novelty and uncertainty boosts. Columns 1–3 interact both with one characteristic of the paper's most prominent author: the highest H-index among its authors is in the top 10% of that year's submissions (column 1), any author is at a top-20 institution (column 2), or any author is senior (column 3). Column 4 includes the three characteristics and their interactions together. Difficulty of evaluation is the referee's uncertainty, from 1 (absolutely certain) to 5 (educated guess). Papers for which a characteristic cannot be measured are kept in the sample with their own indicator, which is also interacted with both measures; this applies to about 1% of papers for each characteristic. Brackets are 95% confidence intervals. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

authors with a high H-index remains significant. Patent citations, available for only a fourteenth of the papers, yield 95% confidence intervals one and a half to almost three times as wide as the estimates obtained with paper citations, rendering them too imprecise for this analysis.

Uncertainty. Papers written entirely by students and postdocs receive no uncertainty boost. That is, when referees and editors are unsure about a paper with a senior author, they give it the benefit of the doubt and overrate it relative to its impact, but they do not do so for papers written by students and postdocs. Authors with a top-10% H-index receive a roughly 40% larger uncertainty boost, while authors at top-20 institutions do not.

Robustness. These results are robust to varying the cutoffs that define a high H-index and a top institution (Table B.8). The disproportionately large novelty boost of authors with a high H-index rises as the cutoff becomes more selective, from 5 percentage points for the top 50% of submissions to 19 percentage points for the top 5%. Similarly, the disproportionately large novelty boost of authors at top institutions lies between 6 and 8 percentage points for every cutoff from the

top 3 to the top 50 institutions. It also holds when universities are ranked using the Times Higher Education or QS rankings of computer science. Another concern is that because the characteristics are measured for the paper's most prominent author, papers with more authors are mechanically more likely to have a prominent one. The results are nearly unchanged when I add fixed effects for the number of authors (Table B.9). In addition, my model may be misspecified if the publication premium ψ varies for certain types of authors. For instance, less established authors may see a larger citation boost from being published at a top conference. To address this, I interact the acceptance indicator with the author characteristics in my regressions. Table B.9 shows that results are virtually identical in magnitude when using this approach.

Finally, these estimates are nearly identical when papers without a citation count are included and their citations set to zero, when reports that do not assess novelty are dropped, when the sample is restricted to papers whose authors' characteristics are known, and when novelty is measured using explicit mentions of the word "novel" (Table B.9).

6 Conclusion

This paper has established two main results about how the peer review process evaluates novel ideas, using the complete record of peer review at a top artificial intelligence conference. First, referees and editors overvalue novel ideas relative to their eventual impact. This result holds whether impact is measured through paper citations, patent citations, or how many times a paper's code is copied on GitHub. This "novelty boost" cannot be explained by novel ideas being more likely to produce big hits, receiving delayed recognition, being more influential to the papers that cite them, or being harder to evaluate.

Second, the novelty boost is entirely absent for papers written by students and postdocs. It is also twice as large when an author's H-index is in the top 10% of that year's submissions, and 77% larger when an author is at a top-20 institution. Interpreted through the lens of my model, these findings cannot reflect accurate statistical discrimination or attitudes towards risk. This suggests that evaluators either hold mistaken beliefs about the quality of novel work from junior authors, or simply have a distaste for ideas from younger researchers.

My results indicate that the editorial process values the novel ideas of junior researchers less than those of established researchers. Given that junior authors often bring fresh ideas and push their fields in new directions, this fact has repercussions not only for the careers of young scientists, but also for the pace and direction of innovation in AI. If new researchers correctly perceive that their novel ideas will be held to a higher standard than those of established authors, then they may be more hesitant to undertake such pathbreaking projects in the first place. The extent to which these patterns generalize to other journals and funding agencies, and whether changes to the design of peer review such as stricter anonymity reduce them, are open questions.

References

ALBERTS, B., M. W. KIRSCHNER, S. TILGHMAN, AND H. VARMUS (2014): "Rescuing US Biomedical Research from Its Systemic Flaws," *Proceedings of the National Academy of Sciences*, 111, 5773–5777.

AMERICAN ACADEMY OF ARTS AND SCIENCES (2008): "ARISE: Advancing Research in Science and Engineering: Investing in Early-Career Scientists and High-Risk, High-Reward Research," Report, American Academy of Arts and Sciences, Cambridge, MA.

ARTS, S., N. MELLUSO, AND R. VEUGELERS (2025): "Beyond Citations: Measuring Novel Scientific Ideas and Their Impact in Publication Text," *Review of Economics and Statistics*, published online January 28, 2025, https://doi.org/10.1162/rest_a_01561.

ASIRVATHAM, H., E. MOKSKI, AND A. SHLEIFER (2026): "GPT as a Measurement Tool," NBER Working Paper 34834, National Bureau of Economic Research.

AYOUBI, C., M. PEZZONI, AND F. VISENTIN (2021): "Does It Pay to Do Novel Science? The Selectivity Patterns in Science Funding," *Science and Public Policy*, 48, 635–648.

AZOULAY, P. AND W. H. GREENBLATT (2025): "Does Peer Review Penalize Scientific Risk Taking? Evidence from NIH Grant Renewals," NBER Working Paper 33495, National Bureau of Economic Research.

BERGER, E. D. (2026): "CSRankings: Computer Science Rankings," https://csrankings.org, accessed November 2026.

BHARADHWAJ, H., D. TURPIN, A. GARG, AND A. ANDERSON (2020): "De-Anonymization of Authors through arXiv Submissions during Double-Blind Review," arXiv preprint arXiv:2007.00177.

BORNMANN, L., A. TEKLES, H. H. ZHANG, AND F. Y. YE (2019): "Do We Measure Novelty When We Analyze Unusual Combinations of Cited References? A Validation Study of Bibliometric Novelty Indicators Based on F1000Prime Data," *Journal of Informetrics*, 13, 100979.

BOUDREAU, K. J., E. C. GUINAN, K. R. LAKHANI, AND C. RIEDL (2016): "Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science," *Management Science*, 62, 2765–2783.

BRABEN, D. W. ET AL. (2014): "We Need More Scientific Mavericks," Letter, *The Guardian*, March 18, 2014, https://www.theguardian.com/science/2014/mar/18/we-need-more-scientific-mavericks.

BROGAARD, J., J. ENGELBERG, AND C. A. PARSONS (2014): "Networks and Productivity: Causal Evidence from Editor Rotations," *Journal of Financial Economics*, 111, 251–270.

CARD, D. AND S. DELLAVIGNA (2020): "What Do Editors Maximize? Evidence from Four Leading Economics Journals," *Review of Economics and Statistics*, 102, 195–217.

CARD, D., S. DELLAVIGNA, P. FUNK, AND N. IRIBERRI (2020): "Are Referees and Editors in Economics Gender Neutral?" *Quarterly Journal of Economics*, 135, 269–327.

COFFMAN, L. C., M. NIEDERLE, AND A. J. WILSON (2017): "A Proposal to Organize and Promote Replications," *American Economic Review*, 107, 41–45.

COLUSSI, T. (2018): "Social Ties in Academia: A Friend Is a Treasure," *Review of Economics and Statistics*, 100, 45–50.

CRISCUOLO, P., L. DAHLANDER, T. GROHSJEAN, AND A. SALTER (2017): "Evaluating Novelty: The Role of Panels in the Selection of R&D Projects," *Academy of Management Journal*, 60, 433–460.

EBLEN, M. K., R. M. WAGNER, D. ROYCHOWDHURY, K. C. PATEL, AND K. PEARSON (2016): "How Criterion Scores Predict the Overall Impact Score and Funding Outcomes for NIH Peer-Reviewed Applications," *PLOS ONE*, 11, e0155060.

EUROPEAN COMMISSION (2005): "Frontier Research: The European Challenge. Report of the High-Level Expert Group on Maximising the Wider Benefits of Competitive Basic Research Funding at European Level," Report EUR 21619, Directorate-General for Research, Luxembourg, https://erc.europa.eu/sites/default/files/2022-09/high-level_expert_group_report_full_report_2005_en.pdf.

FONTANA, M., M. IORI, F. MONTOBBIO, AND R. SINATRA (2020): "New and Atypical Combinations: An Assessment of Novelty and Interdisciplinarity," *Research Policy*, 49, 104063.

GOOGLE SCHOLAR (2025): "Google Scholar Metrics: Top Publications — Artificial Intelligence," https://scholar.google.com/citations?view_op=top_venues&vq=eng_artificialintelligence, accessed October 2026.

HANSEN, A. L. AND S. KAZINNIK (2023): "Can ChatGPT Decipher Fedspeak?" SSRN Working Paper 4399406, https://ssrn.com/abstract=4399406.

HENGEL, E. (2022): "Publishing While Female: Are Women Held to Higher Standards? Evidence from Peer Review," *Economic Journal*, 132, 2951–2991.

HUBER, J., S. INOUA, R. KERSCHBAMER, C. KÖNIG-KERSTING, S. PALAN, AND V. L. SMITH (2022): "Nobel and Novice: Author Prominence Affects Peer Review," *Proceedings of the National Academy of Sciences*, 119, e2205779119.

ICLR (2024): "ICLR 2024 Reviewer Guide," https://iclr.cc/Conferences/2024/ReviewerGuide, accessed October 2026.

IMBENS, G. W. AND C. F. MANSKI (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

JHA, M., J. QIAN, M. WEBER, AND B. YANG (2024): "ChatGPT and Corporate Policies," NBER Working Paper 32161, National Bureau of Economic Research.

KELLY, B., D. PAPANIKOLAOU, A. SERU, AND M. TADDY (2021): "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, 3, 303–320.

LABAND, D. N. AND M. J. PIETTE (1994): "Favoritism versus Search for Good Papers: Empirical Evidence Regarding the Behavior of Journal Editors," *Journal of Political Economy*, 102, 194–203.

LANE, J. N., M. TEPLITSKIY, G. GRAY, H. RANU, M. MENIETTI, E. C. GUINAN, AND K. R. LAKHANI (2022): "Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation," *Management Science*, 68, 4478–4495.

LEAHEY, E., C. M. BECKMAN, AND T. L. STANKO (2017): "Prominent but Less Productive: The Impact of Interdisciplinarity on Scientists' Research," *Administrative Science Quarterly*, 62, 105–139.

LIPTON, Z. C. AND J. STEINHARDT (2019): "Troubling Trends in Machine Learning Scholarship," *ACM Queue*, 17, 45–77.

LOPEZ-LIRA, A. AND Y. TANG (2026): "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models," *Journal of Financial Economics*, 184, 104335.

MARX, M. AND A. FUEGI (2020): "Reliance on Science: Worldwide Front-Page Patent Citations to Scientific Articles," *Strategic Management Journal*, 41, 1572–1594.

——— (2022): "Reliance on Science by Inventors: Hybrid Extraction of In-Text Patent-to-Article Citations," *Journal of Economics & Management Strategy*, 31, 369–392.

MISEROCCHI, F., S. NORAY, AND A. WU (2026): "The Race between Academia and Industry for AI Researchers," Discussion Paper 106/26, RFBerlin.

NATIONAL ACADEMIES (2007): *Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future*, Washington, DC: National Academies Press.

NOSEK, B. A., J. R. SPIES, AND M. MOTYL (2012): "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability," *Perspectives on Psychological Science*, 7, 615–631.

OKIKE, K., K. T. HUG, M. S. KOCHER, AND S. S. LEOPOLD (2016): "Single-Blind vs Double-Blind Peer Review in the Setting of Author Prestige," *JAMA*, 316, 1315–1316.

OPEN SCIENCE COLLABORATION (2015): "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716.

RASTOGI, C., I. STELMAKH, X. SHEN, M. MEILA, F. ECHENIQUE, S. CHAWLA, AND N. B. SHAH (2022): "To ArXiv or Not to ArXiv: A Study Quantifying Pros and Cons of Posting Preprints Online," arXiv preprint arXiv:2203.17259.

SEKAR, K. AND M. E. GALLO (2023): "Advanced Research Projects Agency for Health (ARPA-H): Overview and Selected Issues," CRS Report R47568, Congressional Research Service, https://www.congress.gov/crs-product/R47568.

SERRA-GARCIA, M. AND U. GNEEZY (2021): "Nonreplicable Publications Are Cited More Than Replicable Ones," *Science Advances*, 7, eabd1705.

STEPHAN, P., R. VEUGELERS, AND J. WANG (2017): "Reviewers Are Blinkered by Bibliometrics," *Nature*, 544, 411–412.

TEPLITSKIY, M., H. PENG, A. BLASCO, AND K. R. LAKHANI (2022): "Is Novel Research Worth Doing? Evidence from Peer Review at 49 Journals," *Proceedings of the National Academy of Sciences*, 119, e2118046119.

TOMKINS, A., M. ZHANG, AND W. D. HEAVLIN (2017): "Reviewer Bias in Single- versus Double-Blind Peer Review," *Proceedings of the National Academy of Sciences*, 114, 12708–12713.

UZZI, B., S. MUKHERJEE, M. STRINGER, AND B. JONES (2013): "Atypical Combinations and Scientific Impact," *Science*, 342, 468–472.

VALENZUELA, M., V. HA, AND O. ETZIONI (2015): "Identifying Meaningful Citations," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

WANG, J., R. VEUGELERS, AND P. STEPHAN (2017): "Bias against Novelty in Science: A Cautionary Tale for Users of Bibliometric Indicators," *Research Policy*, 46, 1416–1436.

Appendix Contents

A Model Proofs **25**

A.1 The referee's estimates (Equation 3) . . . 25

A.2 The editor's estimate (Equation 4) . . . 25

A.3 The editor's estimate in terms of referee reports . . . 26

A.4 Impact in terms of observables (Equation 5) . . . 27

A.5 Covariance between the citation error and acceptance . . . 28

A.6 Bounds from regressions with and without acceptance . . . 29

B Additional Results and Robustness **30**

C Estimating the Publication Premium **35**

D Are Referees Really Doing Optimal Shrinkage? **38**

D.1 Model and Predictions . . . 38

D.2 Empirical Evidence . . . 39

A Model Proofs

A.1 The referee's estimates (Equation 3)

Proposition A.1 (Referee's Estimates)**.** *The referee's optimal estimates of novelty and residual quality are* $\hat{n}_R = \mu_n + \lambda_R(n_R - \mu_n)$ *and* $\hat{r}_R = \mu_r + \lambda_r(r_R - \mu_r)$*, where* $\lambda_R = \sigma_n^2/(\sigma_n^2 + \sigma_R^2)$ *and* $\lambda_r = \sigma_r^2/(\sigma_r^2 + \sigma_{\nu_R}^2)$*. The referee's assessment of quality is* $\hat{q}_R = \beta_R \hat{n}_R + \hat{r}_R + \delta_R U$*.*

Proof. By the normal projection formula, conditional on U,

$$\mathbb{E}[n \mid n_R] = \mu_n + \frac{\mathrm{Cov}(n, n_R)}{\mathrm{Var}(n_R)}(n_R - \mu_n) = \mu_n + \frac{\sigma_n^2}{\sigma_n^2 + \sigma_R^2}(n_R - \mu_n) = \mu_n + \lambda_R(n_R - \mu_n),$$

and analogously $\mathbb{E}[r \mid r_R] = \mu_r + \lambda_r(r_R - \mu_r)$. Since $n \perp r$ and the noises of both signals are independent, $\mathbb{E}[q_R \mid n_R, r_R, U] = \beta_R \hat{n}_R + \hat{r}_R + \delta_R U$, because the referee knows U. □

A.2 The editor's estimate (Equation 4)

The editor observes the referee's shrunk estimates of overall quality and novelty $(\hat{n}_R, \hat{q}_R)$, as well as their own signals of novelty and residual quality (n_E, r_E). Their acceptance decision amounts

to estimating quality q_E given these variables.

Proposition A.2 (Editor's Combined Estimate)**.** *The editor's optimal estimate of quality is*

$$\hat{q}_E = \beta_E(1-\theta)\hat{n}_R + \beta_E\theta\, n_E + (1-\alpha)\hat{r}_R + \alpha\, r_E + \delta_E U,$$

where $\theta = \dfrac{\sigma_E^{-2}}{\sigma_n^{-2}+\sigma_R^{-2}+\sigma_E^{-2}}$ *and* $\alpha = \dfrac{\sigma_{\nu_E}^{-2}}{\sigma_r^{-2}+\sigma_{\nu_R}^{-2}+\sigma_{\nu_E}^{-2}}$.

Proof. Since $\hat{q}_R = \beta_R\hat{n}_R + \hat{r}_R + \delta_R U$, conditioning on $(\hat{n}_R, \hat{q}_R, U)$ is equivalent to conditioning on $(\hat{n}_R, \hat{r}_R, U)$, and

$$\begin{aligned}\mathbb{E}[q_E \mid n_E, r_E, \hat{n}_R, \hat{q}_R, U] &= \mathbb{E}[\beta_E n + r + \delta_E U \mid n_E, r_E, \hat{n}_R, \hat{r}_R, U] \\ &= \beta_E\, \mathbb{E}[n \mid n_E, \hat{n}_R] + \mathbb{E}[r \mid r_E, \hat{r}_R] + \delta_E U,\end{aligned}$$

where the second line uses that $(n, n_E, \hat{n}_R) \perp (r, r_E, \hat{r}_R)$ by the independence of n and r and of their noises.

Because $\lambda_R > 0$ and $\hat{n}_R = \mu_n + \lambda_R(n_R - \mu_n)$, conditioning on $\hat{n}_R$ is equivalent to conditioning on n_R. Since $n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, $n_R = n + \varepsilon_R$ and $n_E = n + \varepsilon_E$ with independent normal noises, precision weighting gives

$$\mathbb{E}[n \mid n_R, n_E] = \frac{\sigma_n^{-2}\mu_n + \sigma_R^{-2}n_R + \sigma_E^{-2}n_E}{\sigma_n^{-2}+\sigma_R^{-2}+\sigma_E^{-2}} = \frac{(\sigma_n^{-2}+\sigma_R^{-2})\,\hat{n}_R + \sigma_E^{-2}n_E}{\sigma_n^{-2}+\sigma_R^{-2}+\sigma_E^{-2}}.$$

The second equality uses $\hat{n}_R = (\sigma_n^{-2}\mu_n + \sigma_R^{-2}n_R)/(\sigma_n^{-2}+\sigma_R^{-2})$, which is $\mu_n + \lambda_R(n_R - \mu_n)$ written in terms of precisions. Hence

$$\mathbb{E}[n \mid \hat{n}_R, n_E] = (1-\theta)\hat{n}_R + \theta n_E, \qquad \theta = \frac{\sigma_E^{-2}}{\sigma_n^{-2}+\sigma_R^{-2}+\sigma_E^{-2}}.$$

The same argument applied to r gives $\mathbb{E}[r \mid \hat{r}_R, r_E] = (1-\alpha)\hat{r}_R + \alpha r_E$ with α as stated. Combining the two terms gives

$$\hat{q}_E = \beta_E(1-\theta)\hat{n}_R + \beta_E\theta\, n_E + (1-\alpha)\hat{r}_R + \alpha\, r_E + \delta_E U,$$

which is the expression in the proposition. □

A.3 The editor's estimate in terms of referee reports

Proposition A.3 (Orthogonal Decomposition)**.** *The editor's estimate can be rewritten as*

$$\hat{q}_E = (\beta_E - \beta_R)\hat{n}_R + (\delta_E - \delta_R)U + \hat{q}_R + \zeta,$$

where $\zeta = \beta_E\theta(n_E - \hat{n}_R) + \alpha(r_E - \hat{r}_R)$ *is mean-zero Gaussian and independent of the referee reports* $(\hat{n}_R, \hat{q}_R)$ *given* U*.*

Proof. Starting from Proposition A.2,

$$\begin{aligned}\hat{q}_E &= \beta_E(1-\theta)\hat{n}_R + \beta_E\theta\, n_E + (1-\alpha)\hat{r}_R + \alpha\, r_E + \delta_E U \\ &= \beta_E\hat{n}_R + \hat{r}_R + \delta_E U + \beta_E\theta(n_E - \hat{n}_R) + \alpha(r_E - \hat{r}_R).\end{aligned}$$

Using $\hat{q}_R = \beta_R\hat{n}_R + \hat{r}_R + \delta_R U$, we have $\hat{r}_R = \hat{q}_R - \beta_R\hat{n}_R - \delta_R U$, so

$$\hat{q}_E = \beta_E\hat{n}_R + \delta_E U + (\hat{q}_R - \beta_R\hat{n}_R - \delta_R U) + \zeta = (\beta_E - \beta_R)\hat{n}_R + (\delta_E - \delta_R)U + \hat{q}_R + \zeta.$$

All moments below are conditional on U. ζ is orthogonal to $\hat{n}_R$ because

$$\mathrm{Cov}(n_E - \hat{n}_R, \hat{n}_R) = \mathrm{Cov}(n_E, \hat{n}_R) - \mathrm{Var}(\hat{n}_R) = \lambda_R\sigma_n^2 - \lambda_R\sigma_n^2 = 0,$$

where $\mathrm{Var}(\hat{n}_R) = \lambda_R^2(\sigma_n^2 + \sigma_R^2) = \lambda_R\sigma_n^2$, and $r_E - \hat{r}_R$ is independent of $\hat{n}_R$. By the analogous calculation $\mathrm{Cov}(r_E - \hat{r}_R, \hat{r}_R) = 0$, and $n_E - \hat{n}_R$ is independent of $\hat{r}_R$. Since $\hat{q}_R = \beta_R\hat{n}_R + \hat{r}_R$, both components of ζ are uncorrelated with $\hat{q}_R$ as well, so $\mathrm{Cov}(\zeta, \hat{q}_R) = 0$. Joint normality given U then implies independence of ζ and $(\hat{n}_R, \hat{q}_R)$ given U. If, as assumed in the main text, the variance of ζ does not depend on U, then ζ is also independent of U.

□

A.4 Impact in terms of observables (Equation 5)

Proposition A.4 (Impact Equation)**.** *The impact equation* (2) *can be written as*

$$q_I = (\beta_I - \beta_R)\hat{n}_R + (\delta_I - \delta_R)U + \hat{q}_R + \psi\, A + \nu, \qquad \nu = \beta_I(n - \hat{n}_R) + (r - \hat{r}_R) + \epsilon,$$

where ν *is mean zero and orthogonal to* $(\hat{n}_R, \hat{q}_R, U)$*, regardless of how* μ_n*,* μ_r *and the variances vary with* U*.*

Proof. Adding and subtracting $\beta_I\hat{n}_R$ and $\hat{r}_R$ in $q_I = \beta_I n + \psi A + r + \delta_I U + \epsilon$, and using $\hat{q}_R = \beta_R\hat{n}_R + \hat{r}_R + \delta_R U$,

$$\begin{aligned}q_I &= \beta_I\hat{n}_R + \hat{r}_R + \delta_I U + \psi A + [\beta_I(n - \hat{n}_R) + (r - \hat{r}_R) + \epsilon] \\ &= (\beta_I - \beta_R)\hat{n}_R + (\delta_I - \delta_R)U + \hat{q}_R + \psi\, A + \nu.\end{aligned}$$

$\mathbb{E}[\nu] = 0$ because $\hat{n}_R$ and $\hat{r}_R$ are conditional expectations and $\mathbb{E}[\epsilon] = 0$; in fact $\mathbb{E}[\nu \mid U] = 0$. For

orthogonality conditional on U,

$$\mathrm{Cov}(n - \hat{n}_R, \hat{n}_R) = \mathrm{Cov}(n - \lambda_R n_R, \lambda_R n_R) = \lambda_R \left[\sigma_n^2 - \lambda_R(\sigma_n^2 + \sigma_R^2)\right] = 0,$$

and by the same reasoning $\mathrm{Cov}(r - \hat{r}_R, \hat{r}_R) = 0$; $n - \hat{n}_R$ is independent of $\hat{r}_R$, $r - \hat{r}_R$ of $\hat{n}_R$, and ϵ of all signals. Hence $\mathrm{Cov}(\nu, \hat{n}_R) = \mathrm{Cov}(\nu, \hat{q}_R) = 0$ conditional on U. Since $\mathbb{E}[\nu \mid U] = 0$, the law of total covariance gives $\mathrm{Cov}(\nu, g) = \mathbb{E}[\mathrm{Cov}(\nu, g \mid U)] + \mathrm{Cov}(\mathbb{E}[\nu \mid U], \mathbb{E}[g \mid U]) = 0$ for $g = \hat{n}_R$, $g = \hat{q}_R$ and $g = U$, which is constant given U. How μ_n and μ_r vary with U plays no role. □

A.5 Covariance between the citation error and acceptance

The bounds of Section 4.3 require that $\mathrm{Cov}(A, \nu) \geq 0$: accepted papers have higher impact conditional on referee reports $(\hat{n}_R, \hat{q}_R)$. This subsection gives the condition under which that holds.

Lemma A.1 (Covariance with an indicator)**.** *If (X, Y) is jointly normal, $\mathrm{Var}(Y) > 0$, τ is finite, and $A = \mathbf{1}[Y > \tau]$, then $\mathrm{Cov}(X, A) = z \cdot \mathrm{Cov}(X, Y)$, where $z \equiv \phi\left(\frac{\tau - \mathbb{E}[Y]}{\sigma_Y}\right)/\sigma_Y > 0$ and ϕ is the standard normal density.*

Proof. Write

$$X = \mathbb{E}[X] + \frac{\mathrm{Cov}(X, Y)}{\sigma_Y^2}(Y - \mathbb{E}[Y]) + W,$$

where W is independent of Y. Hence $\mathrm{Cov}(X, A) = \frac{\mathrm{Cov}(X,Y)}{\sigma_Y^2}\mathrm{Cov}(Y, A)$. With $S = (Y - \mathbb{E}[Y])/\sigma_Y$ standard normal and $s = (\tau - \mathbb{E}[Y])/\sigma_Y$, $\mathrm{Cov}(Y, A) = \sigma_Y \mathbb{E}[S\,\mathbf{1}[S > s]] = \sigma_Y \phi(s)$, which gives the result. □

Proposition A.5 (Covariance Between Citation Error and Acceptance)**.** *Conditional on U, the covariance between the citation error ν of Proposition A.4 and acceptance $A = \mathbf{1}[\hat{q}_E > \tau]$ is*

$$\mathrm{Cov}(\nu, A) = z \underbrace{\left[\beta_I \beta_E \theta(1 - \lambda_R)\sigma_n^2 + \alpha(1 - \lambda_r)\sigma_r^2\right]}_{\equiv T}, \tag{13}$$

with $z > 0$ as in Lemma A.1, so that $\mathrm{Cov}(\nu, A) > 0$ if and only if $T > 0$. Let

$$\bar{\beta} \equiv \frac{\alpha(1 - \lambda_r)\sigma_r^2}{\theta(1 - \lambda_R)\sigma_n^2}.$$

Then $T > 0$ whenever $\beta_I \beta_E \geq 0$, and, if $\beta_I \beta_E < 0$, then $T > 0$ if and only if $|\beta_I \beta_E| < \bar{\beta}$.

Proof. By Lemma A.1, $\mathrm{Cov}(\nu, A) = z \cdot \mathrm{Cov}(\nu, \hat{q}_E)$. Recall $\nu = \beta_I(n - \hat{n}_R) + (r - \hat{r}_R) + \epsilon$ and, from Proposition A.2,

$$\hat{q}_E = \beta_E(1 - \theta)\hat{n}_R + \beta_E \theta n_E + (1 - \alpha)\hat{r}_R + \alpha r_E.$$

Since ϵ is independent of every signal, $\mathrm{Cov}(\epsilon, \hat{q}_E) = 0$. The novelty component $n - \hat{n}_R$ is orthogonal to $\hat{n}_R$ (Proposition A.4) and independent of $\hat{r}_R$ and r_E, but

$$\mathrm{Cov}(n - \hat{n}_R, n_E) = \mathrm{Cov}(n, n_E) - \mathrm{Cov}(\hat{n}_R, n_E) = \sigma_n^2 - \lambda_R \sigma_n^2 = (1 - \lambda_R)\sigma_n^2.$$

Similarly, $r - \hat{r}_R$ is orthogonal to $\hat{r}_R$, independent of $\hat{n}_R$ and n_E, and $\mathrm{Cov}(r - \hat{r}_R, r_E) = (1 - \lambda_r)\sigma_r^2$. Combining,

$$\begin{aligned}\mathrm{Cov}(\nu, \hat{q}_E) &= \beta_I \cdot \beta_E \theta \cdot \mathrm{Cov}(n - \hat{n}_R, n_E) + \alpha \cdot \mathrm{Cov}(r - \hat{r}_R, r_E) \\ &= \beta_I \beta_E \theta (1 - \lambda_R)\sigma_n^2 + \alpha(1 - \lambda_r)\sigma_r^2 = T.\end{aligned}$$

The second term of T is always positive, so $T > 0$ when $\beta_I \beta_E \geq 0$; when $\beta_I \beta_E < 0$, $T > 0$ if and only if $|\beta_I \beta_E|\theta(1 - \lambda_R)\sigma_n^2 < \alpha(1 - \lambda_r)\sigma_r^2$, i.e. $|\beta_I \beta_E| < \bar{\beta}$. Without conditioning on U, $\mathrm{Cov}(\nu, A) = \mathbb{E}[\mathrm{Cov}(\nu, A \mid U)] + \mathrm{Cov}(\mathbb{E}[\nu \mid U], \mathbb{E}[A \mid U]) = \mathbb{E}[zT]$, because $\mathbb{E}[\nu \mid U] = 0$ (Proposition A.4). It is therefore positive whenever $T > 0$ at every level of uncertainty, in particular whenever $\beta_I \beta_E \geq 0$. □

A.6 Bounds from regressions with and without acceptance

Proposition A.6 (Bounds on Coefficients). *Consider the regressions of q_I on $(\hat{n}_R, \hat{q}_R)$ as in equation (8), yielding coefficients $(d_1, d_2)'$, and on $(\hat{n}_R, \hat{q}_R, A)$ as in equation (7), yielding coefficients $(c_1, c_2, c_3)'$. When $\psi > 0$ and $\mathrm{Cov}(A, \nu) > 0$:*

(i) d_1 and c_1 weakly bound $\beta_I - \beta_R$ from opposite sides, with ordering determined by the sign of $\beta_E - \beta_R$.

(ii) $d_2 > 1$ and $c_2 < 1$.

(iii) $c_3 > \psi$.

Proof. The impact equation from Proposition A.4 is

$$q_I = (\beta_I - \beta_R)\hat{n}_R + \hat{q}_R + \psi A + \nu.$$

Step 1: Regression without A. Let π be the coefficients on the regression of A on $(\hat{n}_R, \hat{q}_R)$. The omitted-variable-bias formula gives

$$\begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} \beta_I - \beta_R \\ 1 \end{pmatrix} + \psi\pi,$$

using $\mathrm{Cov}(\hat{n}_R, \nu) = \mathrm{Cov}(\hat{q}_R, \nu) = 0$. Notice that this regression of A on $(\hat{n}_R, \hat{q}_R)$ is the OLS analogue of the probit model in equation (9). Under the model's normality assumptions, both sets

of coefficients on the reports $(\hat{n}_R, \hat{q}_R)$ are proportional to $(\beta_E - \beta_R, 1)'$; for OLS, Proposition A.3 and Lemma A.1 give

$$\pi = z \begin{pmatrix} \beta_E - \beta_R \\ 1 \end{pmatrix}, \qquad z > 0.$$

Thus, $d_1 - (\beta_I - \beta_R)$ has the sign of $\beta_E - \beta_R$, and $d_2 > 1$.

Step 2: Regression with A. Let $A^{\perp}$ be the residual from the linear regression of A on $(\hat{n}_R, \hat{q}_R)$, with an intercept, and write $S_A = \mathrm{Var}(A^{\perp}) > 0$. Frisch–Waugh–Lovell gives

$$c_3 - \psi = \frac{\mathrm{Cov}(A^{\perp}, \nu)}{S_A} = \frac{\mathrm{Cov}(A, \nu)}{S_A} > 0,$$

where the second equality again uses that ν is uncorrelated with the reports $(\hat{n}_R, \hat{q}_R)$. Applying the same omitted-variable-bias formula to the regression with A, and using Step 1, gives

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} - c_3 \pi = \begin{pmatrix} \beta_I - \beta_R \\ 1 \end{pmatrix} - (c_3 - \psi)\pi.$$

Since $c_3 - \psi > 0$, the bias $c_1 - (\beta_I - \beta_R)$ has the opposite sign to $\beta_E - \beta_R$, and $c_2 < 1$. Thus, c_1 is the lower bound and d_1 the upper bound when $\beta_E > \beta_R$; their roles reverse when $\beta_E < \beta_R$. If $\beta_E = \beta_R$, both equal $\beta_I - \beta_R$. The same formulas establish the weak bounds when $\psi \geq 0$ and $\mathrm{Cov}(A, \nu) \geq 0$. □

B Additional Results and Robustness

This appendix presents additional results and robustness checks for Section 5.

TABLE B.1: Robustness of Table 2: Alternative Outcomes and Samples

	(1) Baseline	(2) GitHub forks	(3) Patent citations	(4) Uncited kept at zero	(5) Unscored reviews dropped	(6) Pre-decision citations	(7) Mean rating below 5
Novelty	–0.17***	–0.19***	–0.16***	–0.17***	–0.16***	–0.11***	–0.11***
	(0.01)	(0.04)	(0.04)	(0.01)	(0.02)	(0.01)	(0.02)
High uncertainty	–0.37***	–0.25***	–0.13*	–0.37***	–0.38***	–0.25***	–0.19***
	(0.03)	(0.06)	(0.06)	(0.03)	(0.03)	(0.03)	(0.04)
N (reviews)	123,666	42,030	8,704	126,960	85,686	50,270	49,493

Notes: Each column repeats column 2 of Table 2 under an alternative outcome or sample. Column 2: asinh of GitHub forks, for papers with a repository. Column 3: asinh of the number of citing U.S. patents, for 2017–2019 submissions. Column 4 keeps papers without a Google Scholar record, with zero citations. Column 5 drops reviews that do not assess novelty. Column 6: asinh of citations received before the acceptance decision, for papers posted on arXiv before it. Column 7: papers whose average rating is below 5, of which 0.5% are accepted. High uncertainty is relative to very low uncertainty, as in Table 2. Standard errors clustered by paper in parentheses. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

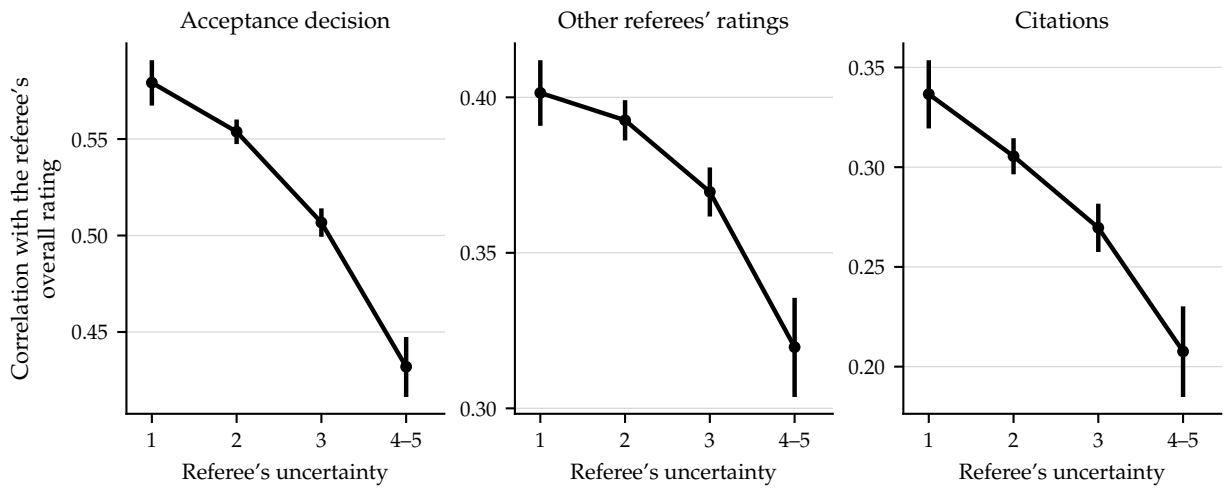


FIGURE B.1: Informativeness of Referee Ratings by Referee Uncertainty

Notes: Correlation of a referee's overall rating with each outcome, by the referee's uncertainty (levels 4 and 5 combined), computed within year with papers weighted equally and combined across years by Fisher's z-transformation; for other referees' ratings, the average correlation with each other referee of the same paper. Bars are 95% bootstrap confidence intervals, resampling papers. The sample is 127,190 reports of 33,082 papers from ICLR 2017–2019 and 2021–2025.

TABLE B.2: NeurIPS 2025 Originality Ratings and Impact

	OLS: asinh citations ($\beta_I - \beta_R$)
	(1)
Originality = 2	−0.10
	(0.08)
Originality = 3	$-0.16^{\dagger}$
	(0.08)
Originality = 4	$-0.17^{\dagger}$
	(0.09)
Rating dummies	Yes
Acceptance control	No
Year and subfield FE	Yes
Referee-count FE	Yes
N	13,762

Notes: This table uses reviews of papers accepted at NeurIPS 2025, whose review form asks referees to rate the originality of each paper from 1 to 4. Each coefficient compares papers given that originality rating with papers rated 1. The regression is otherwise that of column 1 of Table 2, with dummies for the overall rating on the NeurIPS scale of 1 to 6. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE B.3: Top-Cited Indicators: Linear Probability Models

	Top 1% (1)	Top 10% (2)	Top 20% (3)
Novelty	–0.005***	–0.022***	–0.030***
	(0.001)	(0.002)	(0.003)
Relative to base rate	–47%	–22%	–15%
Difficulty of evaluation			
Low	–0.002*	–0.014***	–0.019***
	(0.001)	(0.003)	(0.004)
Medium	–0.007***	–0.035***	–0.045***
	(0.001)	(0.003)	(0.004)
High	–0.007***	–0.053***	–0.075***
	(0.001)	(0.004)	(0.005)
Relative to base rate	–69%	–53%	–38%
Base rate	0.010	0.100	0.199
N	123,666	123,666	123,666

Notes: This table repeats column 4 of Table 2, which controls for acceptance, with an indicator for being highly cited as the outcome: whether the paper's citations are in the top 1, 10 or 20 percent of submissions from the same year. The base rate is the share of papers above each threshold. Standard errors clustered by paper in parentheses. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

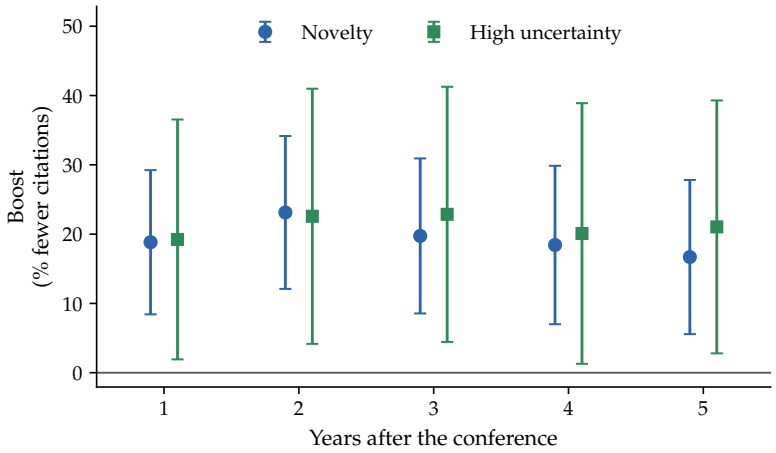


FIGURE B.2: Novelty and Uncertainty Boosts by Years After the Conference

Notes: Each point is a boost: minus 100 times the coefficient on novelty (blue circles) or on high uncertainty, relative to very low uncertainty (green squares), in the regression of column 2 of Table 2, with the inverse hyperbolic sine of the Semantic Scholar citations a paper receives in a given year after the conference as the outcome. The sample is the same papers, submitted from 2017 to 2019, in every year. Bars are 95% confidence intervals clustered by paper.

TABLE B.4: Novelty, Difficulty of Evaluation, and Influential Citations

	All citations (1)	Influential citations (2)
Novelty	–0.21***	–0.18***
	(0.02)	(0.01)
Difficulty of evaluation		
Low	–0.10***	–0.09***
	(0.02)	(0.02)
Medium	–0.28***	–0.23***
	(0.02)	(0.02)
High	–0.44***	–0.33***
	(0.03)	(0.02)
Mean number of citations	90.6	11.6
N	95,051	95,051

Notes: The table repeats column 2 of Table 2 with Semantic Scholar citations as the outcome, for the papers linked to Semantic Scholar. The dependent variable is the inverse hyperbolic sine of the number of citing papers (column 1) or of citing papers whose citation Semantic Scholar classifies as influential (column 2). Standard errors clustered by paper in parentheses. Difficulty of evaluation is relative to very low uncertainty, as in Table 2. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE B.5: Editors' Novelty Boost by Author Characteristics, Without Difficulty of Evaluation

	Editor's novelty boost ($\beta_I - \beta_E$)		
	(1) H-index top 10%	(2) Top-20 institution	(3) Senior
Novelty	-0.14^{***}	-0.11^{***}	0.03
	$[-0.17, -0.11]$	$[-0.15, -0.07]$	$[-0.13, 0.21]$
$\times$ characteristic	-0.13^{**}	-0.07^{**}	-0.18^{*}
	$[-0.24, -0.04]$	$[-0.14, -0.02]$	$[-0.37, -0.02]$
N	123,666	123,666	123,666

Notes: As Table 3, with novelty alone: difficulty of evaluation and its interaction with the characteristic are not included. Brackets are 95% confidence intervals. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE B.6: Reduced-Form Estimates Behind Table 3

	H-index top 10%		Top-20 institution		Senior	
	OLS (1)	Probit (2)	OLS (3)	Probit (4)	OLS (5)	Probit (6)
Novelty	-0.18^{***}	-0.06^{***}	-0.15^{***}	-0.06^{**}	0.01	-0.02
	(0.02)	(0.01)	(0.02)	(0.02)	(0.08)	(0.07)
$\times$ characteristic	-0.13^{**}	0.01	-0.08^{**}	-0.01	-0.21^{*}	-0.04
	(0.05)	(0.04)	(0.03)	(0.02)	(0.08)	(0.07)
N	123,666	123,666	123,666	123,666	123,666	123,666

Notes: The OLS coefficients (asinh citations net of $\hat{\psi}A$) and the acceptance probit coefficients from the joint estimation behind Table B.5, one characteristic at a time. Standard errors in parentheses. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE B.7: Whose Novel Ideas Are Overrated? Referee Novelty Boost by Author Characteristics

	Novelty boost by author characteristic		
	(1) H-index top 10%	(2) Top-20 institution	(3) Senior
Panel A: Editor's novelty boost ($\beta_I - \beta_E$), last author			
Novelty	−0.14***	−0.15***	−0.02
	[−0.17, −0.11]	[−0.18, −0.11]	[−0.15, 0.11]
× characteristic	−0.14**	−0.06†	−0.15*
	[−0.24, −0.05]	[−0.13, 0.00]	[−0.28, −0.02]
Panel B: Referee's novelty boost ($\beta_I - \beta_R$), most prominent author			
Novelty	−0.18***	−0.15***	0.01
	[−0.21, −0.13]	[−0.19, −0.09]	[−0.14, 0.20]
× characteristic	−0.13**	−0.08**	−0.21*
	[−0.23, −0.04]	[−0.15, −0.02]	[−0.38, −0.05]
Panel C: Editor's novelty boost ($\beta_I - \beta_E$), first author			
Novelty	−0.15***	−0.14***	−0.15***
	[−0.19, −0.12]	[−0.17, −0.10]	[−0.18, −0.12]
× characteristic	0.00	−0.06†	−0.02
	[−0.10, 0.09]	[−0.13, 0.00]	[−0.09, 0.06]
N	123,666	123,666	123,666

Notes: Panels A and C repeat Table B.5 for the characteristics of the paper's last and first author. Panel B reports the referee's novelty boost for the most prominent author, from the same estimation. Brackets are 95% confidence intervals. †$p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE B.8: Robustness of Table 3: Alternative Cutoffs and University Rankings

Panel A: Highest H-index among the authors in the top q% of that year's submissions	$q = 5$	$q = 10$	$q = 20$	$q = 30$	$q = 40$	$q = 50$
× characteristic	−0.19**	−0.13**	−0.12***	−0.12***	−0.07*	−0.05†
	[−0.33, −0.06]	[−0.24, −0.04]	[−0.20, −0.05]	[−0.20, −0.06]	[−0.14, −0.01]	[−0.12, 0.00]
Panel B: Any author at one of the top N institutions	$N = 3$	$N = 5$	$N = 10$	$N = 20$	$N = 30$	$N = 50$
Accepted ICLR papers	−0.08*	−0.06†	−0.07*	−0.07**	−0.08**	−0.07*
	[−0.17, 0.00]	[−0.14, 0.01]	[−0.15, −0.01]	[−0.14, −0.02]	[−0.16, −0.03]	[−0.15, −0.01]
Times Higher Education	−0.05	−0.05	−0.05	−0.09**	−0.08**	–
	[−0.16, 0.05]	[−0.15, 0.03]	[−0.14, 0.02]	[−0.16, −0.03]	[−0.15, −0.03]	
QS	−0.02	−0.05	−0.07*	−0.07*	−0.04	–
	[−0.13, 0.07]	[−0.15, 0.03]	[−0.16, −0.01]	[−0.15, −0.02]	[−0.11, 0.01]	
N	123,666	123,666	123,666	123,666	123,666	123,666

Notes: Each cell is the interaction of novelty with one characteristic of the paper's most prominent author, as in Table B.5, defined with another cutoff: the highest H-index among the authors in the top q% of that year's submissions (Panel A; $q = 10$ in Table 3), and any author at one of the top N institutions by accepted ICLR papers or the top N universities in the 2024 Times Higher Education and QS computer-science rankings (Panel B; $N = 20$ by accepted ICLR papers in Table 3). Brackets are 95% confidence intervals. †$p < 0.10$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

TABLE B.9: Robustness of Table 3: Alternative Outcomes, Measures and Samples

	Interaction of novelty with the characteristic			
	(1) H-index top 10%	(2) Top-20 institution	(3) Senior	N (reviews)
Baseline (Table B.5)	−0.13**	−0.07**	−0.18*	123,666
	[−0.24, −0.04]	[−0.14, −0.02]	[−0.37, −0.02]	
GitHub forks	−0.22*	−0.10	−0.16	42,252
	[−0.43, −0.02]	[−0.24, 0.03]	[−0.55, 0.24]	
Keyword-based novelty measure	−0.13*	−0.08*	−0.24*	123,666
	[−0.25, −0.03]	[−0.16, −0.02]	[−0.43, −0.05]	
Publication premium varies by characteristic	−0.13**	−0.07*	−0.18†	123,666
	[−0.22, −0.03]	[−0.13, −0.02]	[−0.32, 0.03]	
Uncited papers kept at zero	−0.13**	−0.08**	−0.17*	126,960
	[−0.24, −0.04]	[−0.15, −0.03]	[−0.36, −0.02]	
Reviews without a novelty assessment dropped	−0.14**	−0.07*	−0.18*	85,686
	[−0.24, −0.04]	[−0.14, −0.02]	[−0.37, −0.02]	
Known characteristics only	−0.13**	−0.07**	−0.18*	122,093
	[−0.24, −0.04]	[−0.14, −0.02]	[−0.37, −0.02]	
Team-size fixed effects	−0.12*	−0.06*	−0.18*	123,666
	[−0.22, −0.03]	[−0.13, −0.01]	[−0.37, −0.03]	

Notes: Each cell is the interaction of novelty with the characteristic in the column, for the paper's most prominent author, as in Table B.5, under the variation in the row. GitHub forks: asinh of forks, for papers with a repository. Keyword-based novelty measure: Online Appendix OA.2. Publication premium varies: the acceptance indicator is also interacted with the characteristic in the upper-bound model. Uncited papers: papers without a Google Scholar record, with zero citations. Known characteristics: papers on which the characteristic is known; N is for the H-index. Team-size fixed effects: one fixed effect for each number of authors from 1 to 6, and one for 7 or more, in both the citation and the acceptance equations. Brackets are 95% confidence intervals. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

C Estimating the Publication Premium

This appendix describes how the publication premium ψ (i.e., the causal effect of acceptance on a paper's citations) is estimated through an event study.

Data Construction. The sample consists of ICLR 2017–2023 submissions posted on arXiv before the decision date whose average referee rating lies between 5 and 6, the range in which both acceptance and rejection are common. The resulting sample contains 2,019 papers, of which 32.4% were accepted.

I build a panel of citations from the paper's Semantic Scholar citation records. Each citing paper is dated at the earlier of its publication date and its arXiv posting date, so that a citation is counted when the citing work first became public. Months are aggregated into bimonthly bins, and the outcome is the inverse hyperbolic sine of the number of citations paper i receives in time t.[24]

Descriptive Statistics. Figure C.1 compares the papers in the sample with all ICLR submissions from 2017 to 2023. The sample papers are accepted at almost exactly the average rate, but they are cited far more than the average submission, by 0.45 standard deviations, and they are

[24]Note that the conference itself takes place 2.5–4.5 months after decisions, depending on the year.

somewhat more likely to have an author with a top-10% H-index or at a top-20 institution, by 0.13 and 0.17 standard deviations, and about as likely to have a senior author.

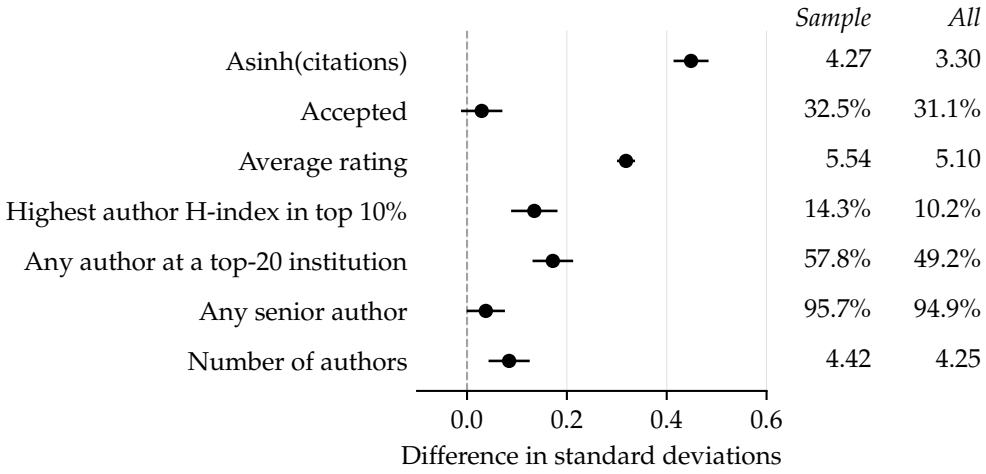


FIGURE C.1: Event-Study Sample Compared with All Submissions

Notes: Each point shows how much the papers in the event-study sample differ from all ICLR submissions from 2017 to 2023. The difference is the mean in the sample minus the mean among all submissions, divided by the standard deviation among all submissions. Bars are 95% confidence intervals. The two columns on the right report the mean in the sample and the mean among all submissions. The author characteristics are defined as in Table 3.

Empirical Strategy. I estimate the equation

$$Y_{it} = \alpha_i + \lambda_t + \sum_{k \neq -1} \beta_k \, \text{Accepted}_i \times \mathbf{1}[t = k] + \varepsilon_{it}, \tag{14}$$

where α_i is a paper fixed effect, λ_t a bimonth fixed effect. The coefficient β_k is the difference between accepted and rejected papers in citations at event time k relative to the last time period before the decision, $k = -1$. Standard errors are clustered by paper.

Even within the papers with a mean referee rating between 5 and 6, accepted papers have somewhat higher ratings and may be cited faster before the decision. I therefore report two versions of every estimate: unweighted, and with rejected papers reweighted by inverse propensity scores, where the propensity score is the predicted probability of acceptance from a logistic regression on indicators for the decile of the average rating, the decile of the pre-decision citation rate, and the submission year.

Results. Table C.1 shows that the two-period estimate of the publication premium is $\hat{\psi} = 0.33$ (s.e. 0.045) reweighted and 0.44 (s.e. 0.037) unweighted. These estimates exclude the period between the decision date and the conference end date. These estimates are also computed on an unbalanced panel, as some papers may only be posted on arXiv one or two months before the decision deadline. Figure C.3 shows results using a balanced panel. The pattern is unchanged when the panel is restricted to papers that were on arXiv for at least 6 or 9 months before the decision.

TABLE C.1: Difference-in-Differences Estimates of the Publication Premium

	Unweighted	IPSW
Publication premium $\hat{\psi}$	0.436***	0.329***
	(0.037)	(0.045)
Pre-decision coefficient	−0.042	0.020
	(0.029)	(0.037)
Papers	2,019	2,019

Notes: The table reports two-period difference-in-differences estimates of equation (14). The IPSW column reweights rejected papers by the inverse of their estimated probability of acceptance. $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

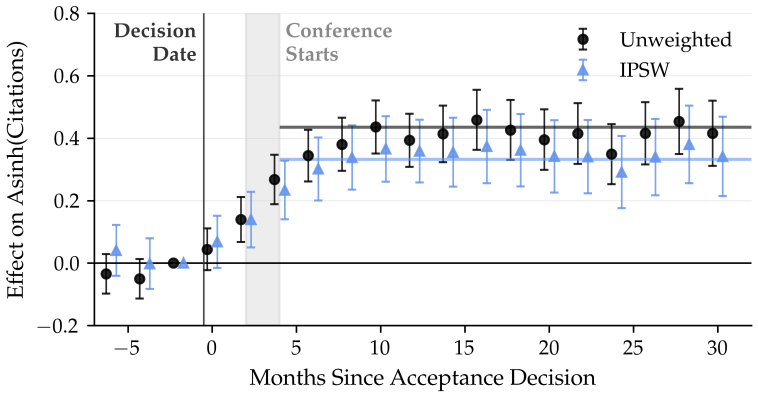


FIGURE C.2: Event Study: Effect of Acceptance on Citations

Notes: The figure shows the estimated coefficients β_k of equation (14).

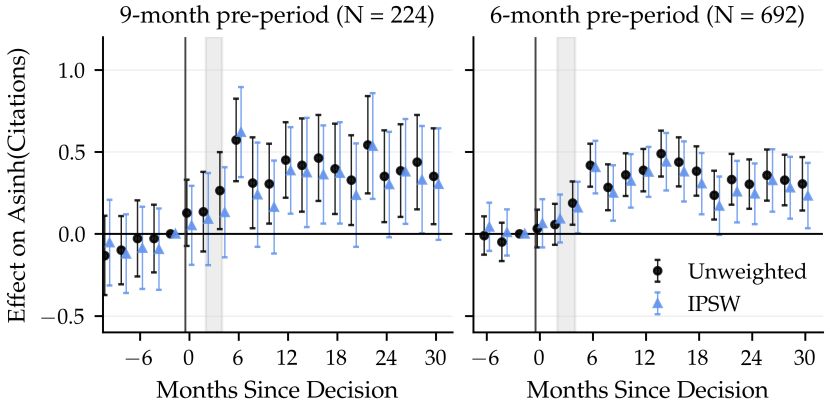


FIGURE C.3: Event Studies on Balanced Pre-Periods

Notes: The figure shows the estimated coefficients β_k of equation (14) for papers posted on arXiv at least 9 months (left) or 6 months (right) before the decision.

D Are Referees Really Doing Optimal Shrinkage?

The model in Section 4 relies on the assumption that referees optimally shrink their estimates of overall quality and novelty when reporting them to editors. This appendix tests that assumption for referees' overall ratings of quality, using referees' uncertainty about their own evaluations as a measure of how precise their information is. I first present a simple model and state its key predictions, and then test each prediction empirically using data from ICLR 2017–2019 and 2021.

D.1 Model and Predictions

Setup. Consider two groups of referees: less uncertain (h, for high precision) and more uncertain (ℓ, for low precision). Each paper has true quality $q \sim N(0, \sigma^2)$, where σ^2 is the dispersion in true quality across papers (the same for both groups). Referees observe noisy signals:

$$S_h = q + \varepsilon_h, \quad \varepsilon_h \sim N(0, \sigma_h^2),$$
$$S_\ell = q + \varepsilon_\ell, \quad \varepsilon_\ell \sim N(0, \sigma_\ell^2),$$

where q, ε_h, and ε_ℓ are mutually independent, $0 < \sigma^2 < \infty$, and less uncertain referees have more precise signals: $0 < \sigma_h^2 < \sigma_\ell^2 < \infty$.

Shrinkage. If referees report Bayesian posterior means (shrinkage estimators), their ratings are:

$$\hat{q}_h = \lambda_h S_h, \quad \hat{q}_\ell = \lambda_\ell S_\ell,$$

where the shrinkage factors are $\lambda_h = \sigma^2/(\sigma^2 + \sigma_h^2)$ and $\lambda_\ell = \sigma^2/(\sigma^2 + \sigma_\ell^2)$. Since $\sigma_h^2 < \sigma_\ell^2$, we have $\lambda_h > \lambda_\ell$: less uncertain referees shrink less.

Key Predictions. The model generates three testable predictions:

Proposition D.1 (Shrinkage in Variance)**.** *When true quality dispersion is the same, less uncertain referees exhibit* ***greater*** *variance in their ratings:*

$$\mathrm{Var}(\hat{q}_h) > \mathrm{Var}(\hat{q}_\ell), \quad \textit{with ratio} \quad \frac{\mathrm{Var}(\hat{q}_h)}{\mathrm{Var}(\hat{q}_\ell)} = \frac{\sigma^2 + \sigma_\ell^2}{\sigma^2 + \sigma_h^2}.$$

Proposition D.2 (Correlation Ratio)**.** *The ratings of less uncertain referees are more strongly correlated with true quality:*

$$\frac{\mathrm{Corr}(\hat{q}_h, q)}{\mathrm{Corr}(\hat{q}_\ell, q)} = \sqrt{\frac{\sigma^2 + \sigma_\ell^2}{\sigma^2 + \sigma_h^2}} > 1.$$

Remark. The ratio of squared correlations (R^2) equals the ratio of variances:

$$\frac{\mathrm{Corr}^2(\hat{q}_h, q)}{\mathrm{Corr}^2(\hat{q}_\ell, q)} = \frac{\mathrm{Var}(\hat{q}_h)}{\mathrm{Var}(\hat{q}_\ell)} = \frac{\lambda_h}{\lambda_\ell}.$$

Proposition D.3 (Characterization of Bayesian Shrinkage). *Consider a referee whose signal is* $S = q + \varepsilon$*, with* $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ *independent of* q*, and assume ratings take the linear form* $\hat{q} = a \cdot S$ *for some nonzero scalar* a*. Then the slope from regressing true quality* q *on ratings* $\hat{q}$ *equals 1* ***if and only if*** $a = \lambda$*:*

$$\frac{\mathrm{Cov}(q, \hat{q})}{\mathrm{Var}(\hat{q})} = 1 \quad \Longleftrightarrow \quad a = \lambda = \frac{\sigma^2}{\sigma^2 + \sigma_\varepsilon^2}.$$

Consequently, the slope equals 1 for all uncertainty levels if and only if each uncertainty level applies its optimal shrinkage factor $\lambda(\sigma_\varepsilon^2)$*. Because citations measure true quality only up to a scalar factor, in practice this implies that the slopes are equal across all uncertainty levels.*

D.2 Empirical Evidence

D.2.1 Test of Proposition D.1: Rating Variance by Uncertainty

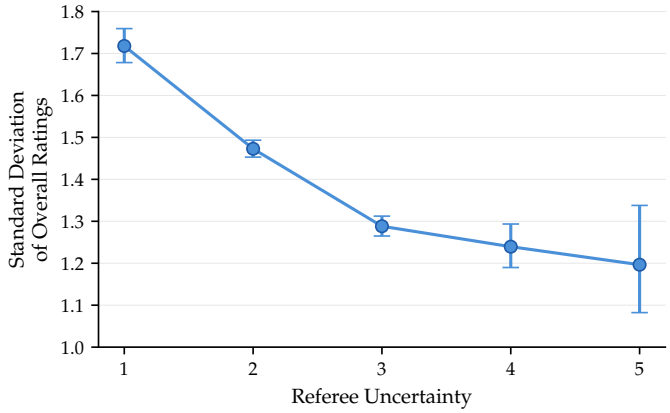


FIGURE D.1: Standard Deviation of Ratings by Referee Uncertainty

Notes: The figure shows the standard deviation of overall ratings (on a scale of 1 to 10) at each level of the referee's uncertainty (1 to 5). Bars show 95% confidence intervals based on the chi-squared distribution. The sample is 18,614 reports from ICLR 2017–2019 and 2021.

D.2.2 Test of Proposition D.2: Correlation with Citations

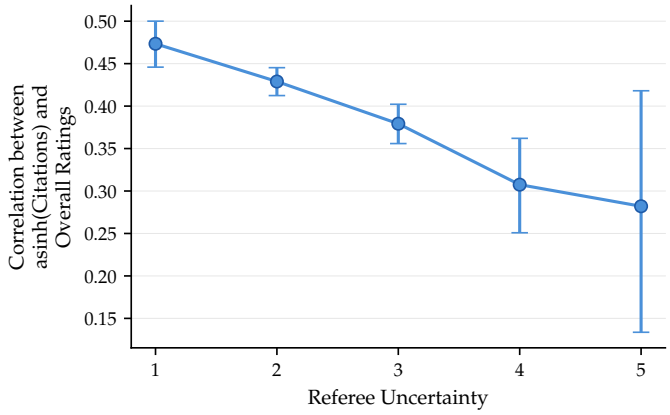


FIGURE D.2: Correlation between Ratings and Citations by Referee Uncertainty

Notes: The figure shows the correlation between a paper's citations (inverse hyperbolic sine) and the referee's overall rating, separately for each uncertainty level. Bars show 95% confidence intervals based on Fisher's z-transformation. The sample is reports of papers with a citation count from ICLR 2017–2019 and 2021.

D.2.3 Test of Proposition D.3: Constant Citation Slopes

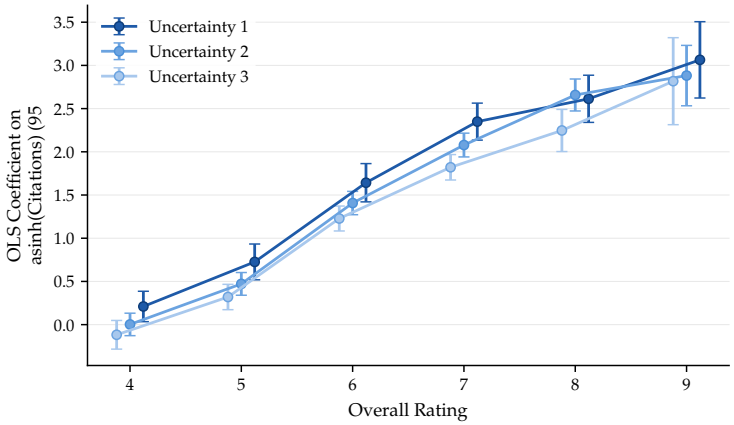


FIGURE D.3: Citation Slopes by Uncertainty Level

Notes: The figure shows the coefficients on rating dummies from regressions of citations (inverse hyperbolic sine), estimated separately for reports with uncertainty 3, 2 and 1; darker shading means lower uncertainty. The regressions include year fixed effects and weight each report by the inverse of the number of referees of its paper. Standard errors are clustered by paper. The sample is ICLR 2017–2019 and 2021.