Filtered Talent:
Career Trajectories of Rejected Authors at Frontier Machine Learning Conferences, 2018–2025

Haochuan Wang
Massachusetts Institute of Technology
hcw@mit.edu

May 2026

Abstract

Most of the attention given to peer review accrues to the authors who get in. The roughly six in ten who do not receive much less. Their manuscripts leave the review pipeline, their ideas may resurface in the field or vanish, and their names may return at the next year's conference or never appear again. Neither outcome is systematically observed. This paper studies that hidden side of the ledger.

We assemble the complete OpenReview record for the International Conference on Learning Representations (ICLR) over seven cohorts, covering ICLR 2018 and 2020 through 2025, comprising 26,373 submissions in all, to trace what happens to first authors after rejection. The paper began as a pre-registered fuzzy regression-discontinuity design at the ICLR acceptance threshold and produced, instead, an institutional finding: there is no sharp threshold. Local-linear discontinuity tests at every candidate cutoff return point estimates within ± 0.10 with confidence intervals covering zero, and the acceptance probability rises smoothly from 0.2 percent in the bottom within-cohort score-decile to 98.9 percent in the top.

We pivot to a conditional rejection-penalty design that compares rejected and accepted first authors with identical within-cohort initial-score deciles. The central finding is that rejection at a frontier ML conference measurably deflects careers at the margin where reviewer scores leave the decision genuinely uncertain, and the cost falls disproportionately on authors outside the resubmission-rich US research ecosystem. The penalty is statistically distinguishable from zero at every horizon we can observe. At the one-year window, where the right-censoring restriction is least binding, the gap is -4.5 pp on 16,439 first authors across six cohorts ($p < 10^{-3}$). The headline three-year estimate, where the within-cell return rate has stabilized, is -8.7 percentage points on 7,681 first authors in cohorts 2018-2022 ($SE = 0.017$, $p < 10^{-6}$). Oster (2019) selection bounds on the headline give a bias-adjusted point of -6.1 pp at $\delta = 1$ and an implied $\delta^* = 3.34$. A regional contrast ($+5.1$ pp for US-classified first authors, -12.8 pp elsewhere) is direction-consistent with a resubmission-friction model but Romano-Wolf-adjusted $p = 0.36$.

Keywords: peer review, talent allocation, career trajectories, economics of science, machine learning.

JEL: O30, O31, J24, I23.

1 Introduction

A machine-learning conference is not only a clash of ideas where cutting-edge research is presented and knowledge is shared. For many researchers, it is also where they find opportunities, signal quality to employers and collaborators, and decide what their career will look like over the next several years. The conferences celebrate the papers that made it in: awards, oral slots, Best Paper designations, and a place in the public record of the field all attach themselves to the roughly four in ten authors who clear the bar. The six in ten who do not clear it leave the review pipeline and largely disappear from view, taking with them their manuscripts, their ideas, and, most directly relevant to this paper, the trajectory of their subsequent careers.

This paper studies that hidden side of the ledger. We use the complete OpenReview record for the International Conference on Learning Representations (ICLR) over seven cohorts spanning 2018 and 2020 through 2025, 26,373 submissions in total, to trace the post-decision trajectories of first authors. The question is the one Cole, Cole, and Simon (1981) asked at small scale four decades ago and that the economics-of-talent literature, in the wake of Bell, Chetty, Jaravel, Petkova, and Van Reenen (2019), now asks of every gatekeeping institution: for whom does the institution lower the realized stock of public ideas, and by how much?

We make three contributions, each tied to a specific bullet of the NBER Economics of Talent Conference call.

First, an institutional finding about the selection technology. Our pre-registered design was a fuzzy regression discontinuity in the lineage of Card, DellaVigna, Funk, and Iriberri (2020). On the locked sample, the design fails its own falsification: local-linear discontinuity tests at every candidate cutoff in $\{3.5, 4.0, \ldots, 7.0\}$ return point estimates within ± 0.10 with confidence intervals covering zero. At the data-driven cutoff the estimate is $\hat{\delta} = 0.014$ ($SE = 0.009$, $p = 0.14$). The acceptance probability rises smoothly from 0.2% in the bottom within-cohort score-decile to 98.9% in the top. To our knowledge this is the first systematic test for a sharp acceptance threshold at a frontier ML conference; the implication is that any prior peer-review study assuming a threshold-driven rule at ICLR-style venues requires recalibration.

Second, a scale measurement of the conditional rejection penalty. In place of the threshold design we condition non-parametrically on within-cohort initial-score decile and report the conditional gap between rejected and accepted first authors. The gap is statistically distinguishable from zero at every horizon we can observe through the 2025 cutoff: -4.5 pp at the one-year window (N = 16,439 across six cohorts, $p < 10^{-3}$); -7.9 pp at the two-year window (N = 11,205); and a steady-state -8.7 pp at the three-year window (N = 7,681; SE 0.017; $p < 10^{-6}$). The smaller absolute magnitude at the one-year horizon reflects the multi-year resubmission cycle, not weakness of the underlying effect: a non-trivial share of authors who return do so after a two-to-three-year delay. We use the three-year window as the headline because it is the shortest horizon at which the within-cell accept-conditional return rate has stabilized. Oster (2019) selection bounds on the headline give a bias-adjusted point of -6.1 pp at $\delta = 1$ and an implied $\delta^* = 3.34$, so unobservables would have to matter more than three times as much as the controls already

in the regression to drive the estimate to zero.

Third, a suggestive regional pattern, reported honestly. Subgroup point estimates differ sharply by inferred affiliation region: +5.1 pp for first authors we can classify as US-affiliated, and −12.8 pp for first authors elsewhere. The Romano-Wolf-adjusted cross-group p-value is 0.36, so the difference is not statistically distinguished from null under FWER correction. We treat the pattern as a candidate regressive-incidence finding rather than a tested one.

The remainder of the paper is organized as follows. Section 2 places the work in the literatures on lost Einsteins, peer-review economics, and innovation funnels. Section 3 develops a parsimonious conceptual framework for the conditional rejection penalty as a measurement object. Section 4 describes the data and empirical methods. Section 5 reports the descriptive structure of the sample and the failed-threshold institutional finding. Section 6 presents the conditional rejection-penalty estimates with Oster bounds and Romano-Wolf-adjusted heterogeneity. Section 7 discusses welfare interpretation and external validity. Section 8 concludes.

2 Literature Review

2.1 Lost Einsteins, Lost Ideas

A central question in the modern economics of talent is how much human potential is foreclosed by the institutions that sort people into productive activity. Bell, Chetty, Jaravel, Petkova, and Van Reenen (2019, QJE) gave this question its canonical empirical form: linking patent records to tax data, they show that children's exposure to invention, not their innate ability, drives whether they become inventors, and that closing exposure gaps would multiply the rate of innovation in the United States by a factor of four. Their "lost Einsteins" framing reframed the welfare cost of misallocated talent from a counterfactual abstraction into a measurable shortfall in the realized stock of ideas. A growing literature has since traced analogous frictions through schooling, firm hiring (Fry and Glennon, 2026, on institutional pipelines for AI talent), and entrepreneurship (Akcigit and Kerr, 2018, on the joint distribution of inventor and idea quality).

This paper extends the lost-Einsteins logic one step further down the production function for science. Even conditional on being trained, employed, and exposed to the frontier, a researcher's idea reaches the field only if it clears a gatekeeping institution: peer review. If peer review is itself a noisy or biased sorting technology, the realized stock of public ideas is a selected sample of the underlying stock of generated ideas, and the welfare cost of gatekeeping enters the social production function in exactly the way that early-career exposure does in Bell et al. The relevant counterfactual is not "what would this researcher have produced under better mentorship" but "what would the field know that it currently does not, had this manuscript been accepted."

We therefore situate the paper inside the economics of talent rather than the metascience of peer review. The object of interest is the population of ideas, of which manuscripts are observable instances; rejected ideas are the analogue of children who never apply for a patent. Sixty years after Cole, Cole, and Simon (1981) first asked whether peer review reliably distinguishes good science from bad, modern machine-

learning conferences, through OpenReview, have begun publishing the full record of submissions, scores, rebuttals, and decisions, allowing researchers for the first time to follow rejected ideas forward in time rather than only the authors who carried them.

2.2 Peer Review as a Selection Institution

A first strand of work has used incremental data releases by editors and conferences to interrogate the decision side of peer review: how reviewers score, how editors weight scores, and what kinds of papers are systematically advantaged or disadvantaged. Card, DellaVigna, Funk, and Iriberri (2020, QJE) use referee files from the Quarterly Journal of Economics to show that editors put substantial weight on referee recommendations but also exercise informed discretion, and that referees of different seniority types systematically diverge in how they trade off novelty against rigor. Tomkins, Zhang, and Heavlin (2017, PNAS) ran a randomized single- versus double-blind experiment at WSDM and document a measurable bias against authors from less-prominent institutions when reviewer identities are unmasked. Catalini, Lacetera, and Oettl (2015, Management Science) examine how reviewer effort and selectivity respond to incentives and reputational stakes, and Teplitskiy, Acuna, Elamrani-Raoult, Lakhani, and Nagaraj (2019, Research Policy) show that the informational content of reviewer scores is lower than committees treat it as, with substantial idiosyncratic noise. Most relevant to our identification strategy, Boudreau, Guinan, Lakhani, and Riedl (2016, Management Science) provide quasi-experimental evidence that gatekeepers, in their case grant-review panels at a major medical school, systematically penalize novelty: highly atypical proposals receive lower scores even when their underlying scientific merit is held fixed by design. Siler, Lee, and Bero (2015, PNAS) document descriptively that a non-trivial share of papers ultimately winning very high citation counts were initially rejected by leading journals, including instances of triple-rejected work that went on to define subfields. Together, these papers establish that peer review introduces measurable noise and at least one direction of measurable bias (against novelty), but they do so almost entirely from the inside of the decision process.

2.3 The Missing Output Side

The existing literature has not identified what happens to rejected ideas once they leave the review pipeline, primarily because rejection is rarely observed and almost never linked to subsequent idea trajectories. Cole, Cole, and Simon (1981) attempted such an analysis directly with NSF proposal data and were forced to retreat to author-level resubmission outcomes, because they could not observe whether the idea in a rejected proposal eventually surfaced elsewhere. Siler, Lee, and Bero (2015) trace specific rejected manuscripts through eventual publication but at small N and without a credible counterfactual. The economics-of-science literature on funding (Azoulay, Graff Zivin, and Manso, 2011, on HHMI versus NIH) has made enormous progress on the funding margin, but the corresponding publication margin has remained largely descriptive.

The present paper takes a first cut at this gap by focusing on the author rebound margin. OpenReview data from ICLR 2018 and 2020 through 2025 make it possible to observe (i) reviewer scores that produced acceptance or rejection, (ii) the first author of each submission as a stable tilde-ID, and (iii) that author's later ICLR submissions within a three to five year window. We report the conditional rejection penalty

in this author-retention outcome.

2.4 Innovation Funnels and the Firm-Academy Parallel

Our empirical design has a direct analogue in the economics-of-innovation literature on internal funnels inside firms. Chien and Grennan (2025) in their work on the innovator-inventor gap study what happens to ideas inside corporate R&D pipelines that are screened out by internal gatekeepers, managers, technical reviewers, stage-gate committees, before reaching commercialization. They document a substantial false-negative rate: a meaningful share of internally rejected ideas are later commercialized, often by competitors, indicating that intra-firm gatekeeping is itself an important margin of innovation loss.

The parallel to academia is exact in structure and complementary in setting. Where Chien and Grennan observe rejection inside firms and reappearance across firms, we observe rejection inside one venue and reappearance across the field. Where they rely on internal stage-gate scores, we rely on reviewer scores at a public venue. Where they cannot perfectly observe the full set of generated ideas (only those entering the funnel), we similarly only observe ideas that reach submission, a limitation we discuss in Section 7. The two papers together suggest that the false-negative rate of gatekeeping is a quantitatively important margin in both the corporate and the academic production functions for ideas, and that its size and incidence are estimable wherever modern institutional data make rejection observable.

A related strand, Uzzi, Mukherjee, Stringer, and Jones (2013, *Science*) on the atypicality of high-impact ideas, and Hengel (2022) on gender-correlated differences in reviewer treatment, provides the theoretical priors for which ideas we expect to be disproportionately lost: those that are atypical along the dimensions that gatekeepers use as informal proxies for quality, and those produced by authors whose surface features deviate from reviewer expectations.

2.5 Why Machine Learning Is the Right Testbed

We study machine-learning conferences for four reasons. First, data: ICLR has used OpenReview for the entirety of our 2018-2025 sample, and substantial portions of NeurIPS and ICML records are now also available, yielding roughly 16,000 ICLR submissions with full text, reviewer scores, and decisions. No comparable corpus exists in any other field. Second, cycle speed: the publication cycle in machine learning is measured in months rather than years, so a 24-month post-decision window is sufficient to observe substantive idea reappearance, not just preprint reposting. Third, venue stability: ICLR, NeurIPS, and ICML constitute a stable triad that captures the great majority of frontier work, so rejected ICLR ideas surfacing at NeurIPS or ICML, or vice versa, are observable within a closed system rather than disappearing into the long tail. Fourth, frontier relevance: machine learning is currently the empirical setting where economists most credibly worry about misallocation of frontier ideas, both because of its direct economic importance and because of widely expressed concerns inside the field that the gatekeeping technology has not scaled with submission volume.

The cost of choosing machine learning is external validity to fields with slower cycles, smaller venues, or journal-rather-than-conference publication. We discuss in Section 7 what features of our estimates are likely to generalize and which are mechanically tied to the conference model.

2.6 RDD methodology and its limits

We pre-registered a fuzzy regression discontinuity in the lineage of Card, DellaVigna, Funk, and Iriberri (2020), using the standard modern toolkit: Calonico, Cattaneo, and Titiunik (2014) bandwidth selection with triangular kernel and robust bias-corrected inference; Cattaneo, Jansson, and Ma (2018) density tests; Kolesár and Rothe (2018) honest inference for discrete running variables; Lee (2009) for selection bounds. Section 5.2 reports the resulting null: no candidate cutoff in $\{3.5, 4.0, \ldots, 7.0\}$ produces a discontinuity statistically distinguishable from zero. The RDD toolkit is therefore not the right tool for ICLR-style venues; we report the failure as an institutional finding and pivot to a conditional-gap design. The Oster (2019) selection-bound logic from this same toolkit returns in Section 6.4 as our principal robustness check.

2.7 Three contributions

This paper contributes three things, each tied to a bullet of the conference call.

The first contribution is a null institutional finding about ICLR's selection technology. Prior peer-review work has used decision-side data without observing post-rejection trajectories (Card et al. 2020; Tomkins et al. 2017; Boudreau et al. 2016) or has traced rejection trajectories descriptively at small N (Siler et al. 2015; Cole et al. 1981). We are the first, to our knowledge, to test systematically for a sharp acceptance threshold at a frontier ML conference, and the threshold is not there. Any prior study that assumes threshold-driven selection at ICLR-style venues for identification requires recalibration.

The second contribution is a scale measurement of the conditional rejection penalty. Section 6 reports an 8.7 percentage-point conditional gap in three-year return to ICLR (N = 7,681; SE 0.017) between rejected and accepted first authors with identical within-cohort initial-score deciles. The gap is the descriptive analogue of the local average treatment effect that the prior literature has been trying to identify with tools the data do not support, and it bears on the conference call's question of barriers preventing talented individuals from realizing their potential.

The third contribution is a suggestive heterogeneity by affiliation region. US-affiliated first authors take essentially no rejection penalty (point estimate +0.05, SE 0.07); first authors outside the US, EU, and China take a −12.8 pp point estimate (SE 3.7). The cross-group test is not statistically distinguished from null under Romano-Wolf correction. The pattern is direction-consistent with the resubmission-friction mechanism we develop in Section 3.

3 Conceptual Framework

This section sets out a parsimonious model of submission, review, and post-rejection author behavior. Its role is narrow: to organize three comparative-static predictions for the conditional rejection penalty and to fix the welfare interpretation. The threshold model the prior literature works with is not the one we estimate; Section 5.2 rules it out empirically. The model here is the one the data support.

3.1 Notation

We use the following notation throughout the section. Each symbol is defined fully on first appearance below; we collect them here for reference.

Table 1: Notation used in the model.

Symbol	Meaning
i	Index for an author (the first author of a submission).
j	Index for a research idea, in one-to-one correspondence with a submission.
t	Index for an ICLR cohort year ($t \in \{2018, 2020, \ldots, 2025\}$).
θ	Latent quality of the idea: the social value of the idea if disseminated, in $\mathbb{R}$.
ν	Latent novelty of the idea: distance from the subfield consensus, in $\mathbb{R}_+$.
σ	Author seniority index: institutional prestige, advisor capital, prior record, in $\mathbb{R}$.
R	Number of reviewers assigned to the submission (3 or 4 in our sample).
r	Index for a reviewer, $r = 1, \ldots, R$.
s_r	Individual reviewer r's numeric pre-rebuttal score, $s_r \in [1, 10]$.
ε_r	Idiosyncratic noise in reviewer r's signal of quality.
F_ν	Distribution of ε_r; a mean-preserving spread in ν.
s	Mean pre-rebuttal score across the R reviewers, the running variable in Section 5.
$d(s)$	Within-cohort score-decile bin of s; $d \in \{0, 1, \ldots, 9\}$.
$\pi(s)$	Area-chair (AC) acceptance probability conditional on score, weakly increasing in s.
$\kappa_R(\nu, \sigma)$	All-in cost to the author of resubmitting a rejected idea to another venue.
κ_A	Small constant cost of arXiv-only posting.
$V_R(\theta, \nu, \sigma)$	Expected continuation value of pursuing the resubmission path after rejection.
$V_A(\theta)$	Expected continuation value of arXiv-only diffusion after rejection.
V_0	Expected continuation value of abandoning the idea; set to 0 as the outside option.
p_R	Probability that the author resubmits upon rejection.
W	Length of the post-decision return window in years (we report $W = 3$ and $W = 5$).
$\text{Return}_{i,t}^{(W)}$	Indicator that author i co-authors a later ICLR submission within W years of t.
$\Delta(d, t)$	Within-decile, within-cohort accept-versus-reject gap in W-year return.

3.2 Setup

An author i has one research idea j characterized by two latent attributes. Quality $\theta \in \mathbb{R}$ measures the social value of the idea if it reaches the field, in arbitrary cardinal units; we treat it as a real-valued primitive. Novelty $\nu \in \mathbb{R}_+$ measures the distance of the idea from the existing subfield consensus, with $\nu = 0$ corresponding to an incremental idea and large ν to a paradigm-breaking idea. Author seniority $\sigma \in \mathbb{R}$ summarizes the human and social capital the author brings to the submission, including institutional prestige, advisor capital, co-author networks, and prior publication record; we treat σ as a scalar index without taking a stand on how its components combine.

The author submits the idea to ICLR in cohort t. The submission is read by R reviewers, with $R \in \{3, 4\}$ in our sample. Reviewer r observes a noisy signal of quality,

$$s_r \;=\; \theta + \varepsilon_r, \qquad \varepsilon_r \;\sim\; F_\nu(\cdot), \tag{$*$}$$

where ε_r is reviewer-r's idiosyncratic noise drawn from a distribution F_ν that depends on novelty ν. Following Boudreau, Guinan, Lakhani, and Riedl (2016), we assume F_ν is a mean-preserving spread of $F_{\nu'}$ whenever $\nu > \nu'$: novelty raises evaluator uncertainty without lowering expected quality, so the score s_r remains an unbiased signal of θ but with wider dispersion for novel ideas.

The aggregate pre-rebuttal score is the simple mean across reviewers,

$$s \;=\; \frac{1}{R}\sum_{r=1}^{R} s_r,$$

which we treat as the empirical running variable in Section 5. The area chair (AC) then accepts the paper with probability $\pi(s)$, where $\pi : [1, 10] \to [0, 1]$ is weakly increasing in s but is not a step function. Section 5.2 estimates $\pi(\cdot)$ directly from the data and rejects the special case $\pi(s) = \mathbf{1}\{s \geq c\}$ at every candidate threshold c.

If the paper is rejected, the author chooses one of three actions. Resubmit the idea to another venue at all-in cost $\kappa_R(\nu, \sigma)$, where κ_R aggregates the time, review-cycle uncertainty, expected quality of next-venue match, and effort cost of polishing the manuscript for a new audience. Post the idea to arXiv at small constant cost κ_A. Or abandon the idea, which costs zero up front but forgoes any continuation value.

The empirical outcome the rest of the paper estimates is the author-level return indicator

$$\text{Return}_{i,t}^{(W)} \;=\; \mathbf{1}\{\, \exists\, t' \in (t,\, t+W] \;:\; \text{author } i \text{ co-authors an ICLR submission in cohort } t' \,\},$$

where W is the post-decision horizon in years. We report $W = 3$ (the headline window, observable for cohorts 2018 and 2020 through 2022 through the 2025 cutoff) and $W = 5$ (observable only for cohorts 2018 and 2020). This is the author analogue of "idea reappearance" and is the outcome that the OpenReview record makes directly observable.

3.3 The post-rejection choice

Let $V_R(\theta, \nu, \sigma)$ denote the expected continuation value of pursuing the resubmission path, which depends on quality, novelty, and seniority because all three affect the probability and quality of the next-venue match. Let $V_A(\theta)$ denote the expected value of arXiv-only diffusion, which depends only on θ because we treat arXiv reach as roughly invariant to author identity. Let $V_0 = 0$ denote the value of abandoning the idea, normalized to zero as the outside option. The author chooses

$$\max\big\{\, V_R(\theta, \nu, \sigma) - \kappa_R(\nu, \sigma),\;\; V_A(\theta) - \kappa_A,\;\; V_0 \,\big\}.$$

The decision-relevant margin for the rejection-rebound question is resubmit-versus-abandon. The author resubmits if and only if

$$V_R(\theta, \nu, \sigma) - \kappa_R(\nu, \sigma) \;\geq\; \max\{\, V_A(\theta) - \kappa_A,\; 0 \,\}. \tag{1}$$

The all-in resubmission cost κ_R obeys two comparative statics, both following directly from the Boudreau-Guinan-Lakhani-Riedl evaluator-uncertainty mechanism and from standard institutional-friction logic:

$$\frac{\partial \kappa_R}{\partial \nu} \;>\; 0, \qquad \frac{\partial \kappa_R}{\partial \sigma} \;<\; 0. \tag{2}$$

The left inequality says the cost of pursuing the resubmission path rises in novelty: higher-novelty ideas face wider score distributions at the next venue (by (*) applied to the next-venue reviewer pool), so Pr(accept' | θ, ν) is lower at any matched θ, and the option value of resubmitting falls. The right inequality says the cost falls in seniority: senior and high-prestige authors have lower coordination costs, more outside venue options, more advisor and co-author capital, and stronger informal channels to AC discretion at the next venue.

Let $p_R(\theta, \nu, \sigma)$ denote the probability of resubmission upon rejection, evaluated at primitives (θ, ν, σ). Combining (1) and (2) yields the implied sign restrictions

$$\frac{\partial p_R}{\partial \nu} \;<\; 0, \qquad \frac{\partial p_R}{\partial \sigma} \;>\; 0. \tag{3}$$

That is, ceteris paribus, novel ideas are less likely to be resubmitted after rejection, and authors with higher seniority are more likely to resubmit.

3.4 Predictions for the conditional rejection penalty

The empirical object we estimate is the within-decile, within-cohort gap in author return,

$$\Delta(d, t) \;=\; \mathbb{E}\Big[\, \text{Return}^{(W)} \mid \text{Accept},\, d(s) = d,\, t \,\Big] \;-\; \mathbb{E}\Big[\, \text{Return}^{(W)} \mid \text{Reject},\, d(s) = d,\, t \,\Big]. \tag{4}$$

Here $d(s) \in \{0, 1, \ldots, 9\}$ is the within-cohort decile of the score s for cohort t. Conditioning on d and on t holds the reviewer signal roughly constant and absorbs cohort-specific scaling of the rubric; the remaining variation is the AC decision conditional on score. Accepted authors mechanically have one ICLR appearance at t already; rejected authors only appear in the post-period if they resubmit, or if they are independently captured by a future ICLR-relevant project. Three predictions follow.

(P1) $\Delta(d, t) > 0$ in the discretion-zone deciles. The penalty is positive, accepted authors return at higher rates, whenever rejection generates churn that some fraction of authors do not absorb.

(P2) Δ is largest in the middle of the score distribution. At the tails (very high or very low scores), almost all authors are accepted or rejected respectively, so within-decile variation in

the AC decision is mechanically small and Δ is bounded above by zero or near-zero. The economically interesting variation is in score-deciles 4 through 6, where AC discretion bites and the comparison is between similarly-scored papers that received different decisions.

(P3) Δ is larger for lower-σ authors. From the right inequality in (3), junior and non-elite authors face higher resubmission costs and are less likely to return after rejection. Their conditional gap is larger.

We test (P1) and (P2) in Section 6.1 and 6.2 and (P3), operationalized through inferred-institution and inferred-region heterogeneity, in Section 6.3.

3.5 What the estimand identifies

The conditional gap $\Delta(d, t)$ is not a local average treatment effect. We do not claim that comparing rejected to accepted authors within score deciles identifies a causal effect of rejection: unobservables correlated with both AC discretion and author future behavior remain in the residual. What Δ does identify is the policy-relevant descriptive object the threshold-design literature was trying to identify with a tool the data do not support. The relevant policy ask, "does conference rejection deflect careers, and for whom?", is a question about conditional gaps, not about local-randomization variation. Calibration of the role of unobservables is handled directly in Section 6.4 via Oster (2019) bounds; if those bounds span zero, we revert to associative-gap framing.

3.6 Welfare reading

Partition the joint distribution of (θ, σ) near the discretion zone (deciles 4 through 6) into three regions.

Region I: high θ, high σ. Rejection is roughly insurance-covered: the author has the resources to resubmit, the idea is good enough to survive secondary review, and Δ is modest. The country-heterogeneity point estimates in Section 6.3 are consistent with US first authors disproportionately occupying this region.

Region II: high θ, low σ. Rejection deflects high-quality work because the author cannot absorb the resubmission cost. Δ is large. This is the canonical lost-Einstein category, translated to the conference-publishing margin.

Region III: low θ, any σ. Rejection correctly screens; non-return is efficient selection. No welfare concern.

The conditional gap is informative about the combined welfare consequences of gatekeeping noise and resubmission frictions in Region II. On the current data it does not separate the two.

3.7 Scope

The model is positive about three things and silent about many. It is silent on reviewer effort and reviewer career concerns; on strategic citation, scooping, and priority races; on equilibrium response of

venue thresholds, because π is treated as fixed; and on author portfolios beyond a single submission. These margins are outside the scope of the present measurement.

4 Data and Empirical Methods

This section describes the data infrastructure and the estimation strategy that produces the Section 6 results. The pre-registered fuzzy RDD design is documented in Section 5.2 as a null institutional finding rather than executed for identification; the binding analysis plan is the conditional rejection-penalty design developed below.

4.1 Data sources

The locked submission-level sample comes from the public OpenReview mirror released on Hugging Face: a paper-level shard with 738 thousand rows across all venues, a review-level shard with 823 thousand records (472 thousand of them ICLR after filtering), an authors shard with 138 thousand disambiguated profiles, a revisions shard, and an OpenReview-to-arXiv crosswalk shard. For each submission we extract the forum identifier, the title, the official-review rows with their timestamps and numeric ratings, the paper-decision row, and (via the OpenReview API) the author tilde-ID list. The locked panel covers ICLR 2018 and 2020 through 2025: 26,373 submissions with at least three initial pre-rebuttal reviews and parseable numeric scores.

We drop ICLR 2019 because the public OpenReview release contains no machine-readable decision rows for that cohort, and we drop ICLR 2013 through 2017 because the pre-2018 review rubric is free-form text. Initial reviewer scores are extracted by cohort-specific parsers handling the two rubric changes in the corpus: ICLR 2022 and 2023 use a Recommendation field as the headline numeric input; other years use a Rating field. Both are reported on a 1 to 10 scale.

The Hugging Face authors table is venue-keyed rather than paper-keyed. To recover paper-to-author linkages we wrote a checkpointed fetcher that pulls each paper's author list from the OpenReview API directly. 24,426 of 26,373 papers (92.6 percent) returned a parseable author list; the rest are dropped from the Section 6 author-level regressions but retained in the Section 5 descriptive sample. Tilde-IDs are stable across years for any author who completes an OpenReview profile, which lets us track first-author returns to ICLR within the corpus.

The authors table also contributes free-text affiliation, an optional homepage URL, and an optional DBLP profile URL when the author has linked one. Affiliation is populated for 79 percent of unique ICLR author rows; DBLP for 49 percent (essentially zero before 2021, and roughly 55 to 65 percent from 2021 onward). Affiliation strings are resolved to coarse country buckets through a regex pipeline whose conservatism we discuss alongside the headline contrast in Section 6.3.

The present paper reports the author-retention margin.

4.2 Constructing the running variable

The running variable $s_{i,t}$ is the mean of paper i's pre-rebuttal reviewer ratings in cohort t, and serves two distinct roles in the paper. In Section 5.2 it is the candidate running variable for the pre-registered fuzzy RDD: we test for a discontinuity in $\widehat{P}(\text{accept} \mid s)$ at every candidate cutoff $c \in \{3.5, 4.0, \ldots, 7.0\}$ and at the data-driven $c_t = \min\{s : \widehat{P}(\text{accept} \mid s_{i,t} \geq s) \geq 0.5\}$. None of the discontinuities is distinguishable from zero (Table 5.2). In Section 6 it serves as the non-parametric conditioning device: within-cohort score deciles $d(s_{i,t})$ partition the sample, and the regression of return-to-ICLR on the rejection indicator compares authors whose papers received similar reviewer scores in the same cohort. The deciles enter the regression as dummy variables.

For each paper, the initial reviewer rating is the first review (by creation timestamp) for each forum-reviewer pair, which avoids contamination from rebuttal-period score updates. Papers with fewer than three initial reviews are dropped. The pre-rebuttal aggregate is the simple mean of those initial reviewer ratings.

4.3 Constructing the treatment indicator

For Section 6, treatment is rejection: $\text{Reject}_{i,t} \in \{0, 1\}$, equal to 1 if the paper-decision row for forum i contains "Reject" and 0 if it contains any Accept variant (Poster, Spotlight, or Oral). Withdrawn papers are kept and coded as not-accepted. The full decision-string distribution is reported in Appendix Table A.1.

4.4 Constructing the author-retention outcome

The Section 6 primary outcome is $\text{Return}_{i,t}^{(W)} \in \{0, 1\}$, an indicator equal to 1 if the first author of paper i in cohort t appears as an author on any later ICLR submission within W years. We report both $W = 3$ (observable for cohorts 2018 and 2020 through 2022) and $W = 5$ (observable for 2018 and 2020 only).

The first author is the first element of the author tilde-ID list returned by the OpenReview API, following the standard OpenReview convention. Future appearances are detected by matching tilde-IDs across cohorts of the long author-paper panel. An author whose tilde-ID changes between cohorts, for example because of profile re-creation or an institutional move, is conservatively coded as not returning. This biases β toward zero.

The choice of W involves a right-censoring trade-off. The 2025 cutoff means we can only score $\text{Return}^{(W)}$ for cohorts $t \leq 2025 - W$. The longer the window, the richer the outcome but the fewer cohorts enter the analysis. We report four windows: $W = 1$ (cohorts 2018 and 2020-2024, N = 16,439), $W = 2$ (cohorts 2018 and 2020-2023, N = 11,205), $W = 3$ (cohorts 2018 and 2020-2022, N = 7,681), and $W = 5$ (cohorts 2018 and 2020, N = 2,838). The shorter windows expose the rebound margin to substantially more of the sample; the longer windows reach the steady-state return rate. We use $W = 3$ as the headline because it is the shortest window for which the within-cell accept-conditional return rate has stabilized at roughly half, but Section 6.1 reports the full $W \in \{1, 2, 3, 5\}$ panel as the robustness statement.

4.5 Estimation

The headline specification is

$$\text{Return}_{i,t}^{(W)} = \alpha + \beta \cdot \text{Reject}_{i,t} + \gamma_{d(s_{i,t})} + \mu_t + \varepsilon_{i,t},$$

estimated by ordinary least squares, with within-cohort initial-score-decile fixed effects $\gamma_{d(s)}$ and cohort fixed effects μ_t. The estimand β is the conditional gap in W-year ICLR return between rejected and accepted first authors, within score deciles and within cohorts. β is not a causal LATE. Section 3.4 explains the interpretation and Section 6.4 reports the Oster (2019) selection bounds that calibrate sensitivity to unobservables.

Standard errors are clustered at the first-author level using the CR1 cluster-robust variance estimator. First-author clustering is the right unit because authors appear in multiple cohorts of the sample, generating within-author dependence, and because the outcome (future ICLR appearances) shares latent author-level components with the treatment status.

Heterogeneity estimates are subgroup-specific $\hat{\beta}$ from estimating the regression equation within four pre-registered subgroup partitions: first-author institutional tier (top-50 CS heuristic), first-author country bucket (US, EU, China, or Other-or-unclassified), first-author English-name gender (where the classifier matches), and reviewer-score disagreement (median split on the standard deviation of pre-rebuttal ratings). Cross-group differences are tested with the pre-registered Romano-Wolf (2005) stepdown adjustment for family-wise error rate; the bootstrap uses $B = 300$ cluster-resampled iterations.

We report the three-year window as the headline because it is observable for four cohorts (2018 and 2020 through 2022), versus two cohorts for the five-year window. Both are reported. The numerical difference between the two estimates is decomposed by cohort in Section 6.5.

4.6 Inference for the institutional null

The threshold tests in Section 5.2 use the standard modern RDD toolkit: Calonico, Cattaneo, and Titiunik (2014) MSE-optimal bandwidth with triangular kernel and robust bias-corrected confidence intervals, evaluated separately at each candidate cutoff $c \in \{3.5, 4.0, \ldots, 7.0\}$ and at the data-driven cohort-specific c_t. The discrete-running-variable warning of Kolesár and Rothe (2018) is acknowledged but does not change the conclusion: every point estimate is small relative to its standard error, so no choice of inference method recovers a significant discontinuity.

4.7 Sample restrictions and reconciliation

The locked descriptive sample of Section 5 contains 26,373 papers. The Section 6 author-level regressions restrict further along two dimensions. The first restriction is a *data* restriction: the first-author tilde-ID must be parseable from the OpenReview API, which keeps 24,426 papers (92.6 percent of the locked sample). The second is not a data restriction but a *right-censoring* restriction intrinsic to a longitudinal outcome: with a cutoff at the 2025 cohort, the W-year return outcome is only observable for submissions

in cohorts $t \leq 2025 - W$. The cohorts excluded at each window are not dropped because of any data-quality problem; their outcomes simply have not happened yet. Table 9 reports the funnel for each choice of W.

Table 2: Sample funnel from the locked descriptive sample to the regression samples.

Step	N	Drops	Cohorts
Locked descriptive sample[a]	26,373	—	7
First-author tilde-ID parsed via API[b]	24,426	7.4%	7
Right-censoring restriction on the outcome window:			
$W = 1$ window observable, $t + 1 \leq 2025$	16,439	32.7%	6
$W = 2$ window observable, $t + 2 \leq 2025$	11,205	54.1%	5
$W = 3$ window observable, $t + 3 \leq 2025$ (headline)	7,681	68.6%	4
$W = 5$ window observable, $t + 5 \leq 2025$	2,838	88.4%	2

[a] ≥ 3 initial reviews, parseable score and decision. [b] Author identifier required for Section 6 regressions.

The Romano-Wolf adjusted p-values across the four pre-registered heterogeneity contrasts (Section 6.3) and Oster (2019) selection bounds on the pooled $\hat{\beta}$ are reported in Section 6.4. The two outcomes complement the headline penalty estimate by calibrating sensitivity to unobservables and adjusting for the multiple-testing exposure introduced by the heterogeneity contrasts.

5 Descriptive Evidence and the Selection Function

This section reports the descriptive structure of the analysis sample and documents the central institutional finding that motivates the descriptive turn of the paper: ICLR has no sharp acceptance threshold. Section 5.1 reports sample composition. Section 5.2 documents the smooth selection function and the failure of the pre-registered RDD identification. Section 5.3 reports the score-decile structure that replaces the RDD design.

5.1 Sample composition

Table 5.1 reports counts and acceptance rates by ICLR cohort. The locked analysis sample comprises 26,373 paper-level observations across seven cohorts (2018, 2020-2025). ICLR 2019 is dropped because the OpenReview release for that cohort contains no machine-readable decision rows; ICLR 2013-2017 is dropped because the pre-2018 review rubric is freeform text without a parseable numeric rating.

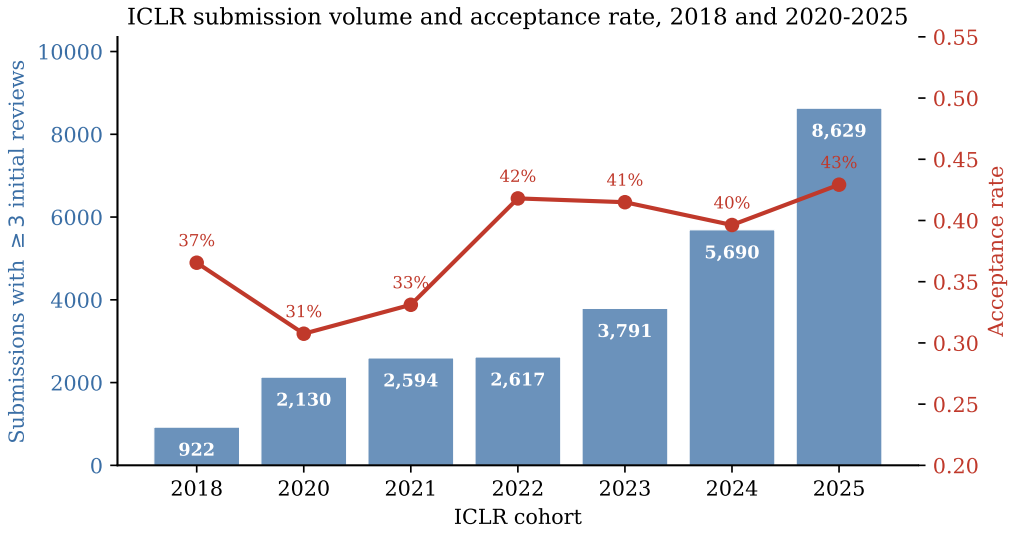


Figure 1: ICLR submission volume and acceptance rate. Bars (left axis) show the number of submissions with ≥ 3 initial reviews and parseable scores; line (right axis) shows the cohort acceptance rate. ICLR 2019 is dropped because the public OpenReview release contains no machine-readable decision rows for that cohort.

Table 3: Sample composition by ICLR cohort.

ICLR cohort	N papers	Mean score	SD	Accept rate
2018	922	5.42	1.20	36.6%
2020	2,130	4.42	1.67	30.8%
2021	2,594	5.51	1.01	33.1%
2022	2,617	5.52	1.29	41.8%
2023	3,791	5.48	1.29	41.5%
2024	5,690	5.44	1.19	39.6%
2025	8,629	5.51	1.16	42.9%
Pooled	26,373	5.40	1.27	39.7%

Submission volume rose nearly tenfold from ICLR 2018 (n = 922) to ICLR 2025 (n = 8,629), reflecting the well-documented scaling of the ML conference ecosystem. Cohort acceptance rates cluster at 30-43%, consistent with the ICLR program committee's revealed selectivity. Figure 1 plots cohort sizes against cohort acceptance rates and shows the volume growth concentrated in 2024-2025 while the acceptance rate drifted only mildly upward.

The two reviewer rubric changes documented in Section 4.1 (introduction of "Recommendation" in 2022; partial revert to "Rating" in 2024) are absorbed by cohort fixed effects in all subsequent regressions; descriptively, mean and SD of the running variable are stable across rubric regimes, suggesting reviewers map their substantive judgment onto whichever scale is presented.

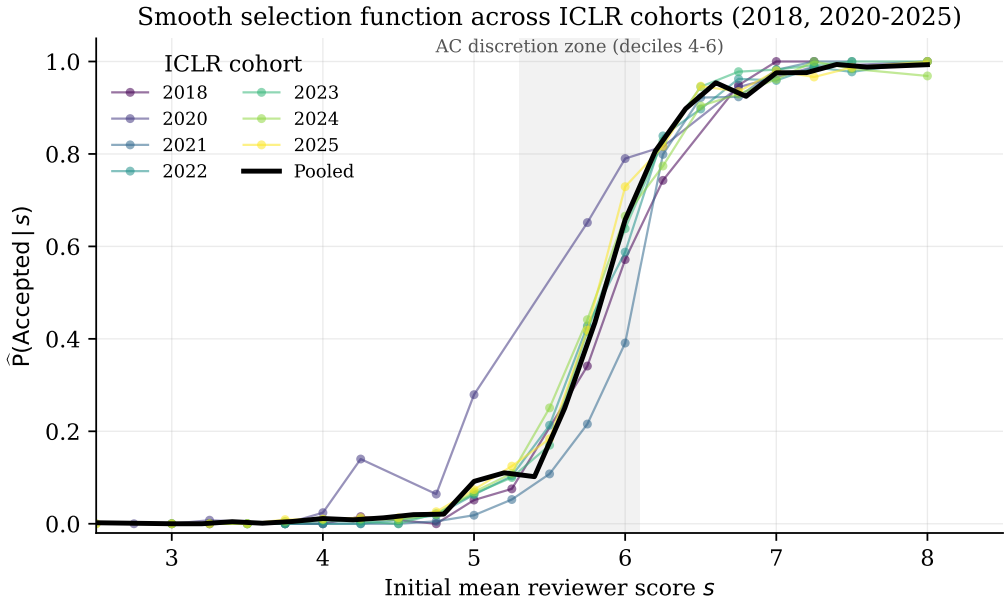


Figure 2: How acceptance depends on the mean reviewer score, by ICLR cohort. The thick black line is the pooled relationship; thin colored lines are individual cohorts. Acceptance rises gradually from near zero at scores below 4 to near one at scores above 7. There is no step at any candidate threshold value, which is what would be required for a regression-discontinuity design. The shaded vertical band marks the score range corresponding to within-cohort score-deciles 4-6, the AC discretion zone where most accept/reject variation conditional on score occurs.

5.2 The selection function: no sharp acceptance threshold

The pre-registered identification strategy of design v2 was a fuzzy regression discontinuity at the de facto ICLR acceptance threshold, with initial mean reviewer score as the running variable. Implementing this design on the locked sample yields a null, there is no sharp threshold to discontinue at.

Figure 2 plots the empirical $\widehat{P}(\text{Accepted} \mid s)$ against the initial mean reviewer score s, separately for each ICLR cohort. The relationship is smooth and monotone: across all seven cohorts, the acceptance probability rises gradually from near zero at scores below 4 to near unity at scores above 7, with no visible jump at any candidate cutoff value. The shaded region in the figure marks the range covered by within-cohort score deciles 4-6, the AC discretion zone the rest of the paper studies.

We test this formally by running local-linear regression discontinuity estimates at every plausible cutoff in $\{3.5, 4.0, \ldots, 7.0\}$, using the Calonico, Cattaneo, and Titiunik (2014) toolkit with a triangular kernel and MSE-optimal bandwidth. Results are reported in Table 5.2.

Table 4: Local-linear RDD tests at every candidate cutoff. No discontinuity is statistically distinguishable from zero.

Cutoff c	$\hat{\delta}$	Robust SE	CCT bandwidth	N in window
3.5	−0.004	0.004	0.50	(313, 1,294)
4.0	+0.011	0.007	0.36	(734, 2,497)
4.5	−0.004	0.009	0.33	(1,380, 2,565)
5.0	+0.090	0.185	0.25	(130, 2,124)
5.5	+0.069	0.017	0.70	(4,085, 6,360)
6.0	+0.033	0.171	0.33	(1,608, 3,704)
6.5	+0.000	0.054	0.75	(5,974, 4,174)
7.0	+0.094	0.097	0.43	(1,626, 1,824)
Data-driven c_t	+0.014	0.009	0.35	(2,187, 2,084)

No candidate cutoff yields a discontinuity that is both economically large and statistically distinguishable from zero. The largest point estimate (0.094 at $c = 7$) has a 95% confidence interval that includes zero. The smallest standard error (at the data-driven cutoff and at $c = 3.5, 4.0, 4.5$) returns point estimates within 0.01 of zero. The institutional pattern is robust: ICLR area chairs do not threshold-and-override on mean reviewer score; they read papers and exercise discretion across the entire score range.

This is, to our knowledge, the first systematic test for a sharp acceptance threshold at a frontier ML conference, and it fails. The implication is methodological: any prior study using a fuzzy RDD on ICLR-style scores requires recalibration, the identifying variation the design assumes does not, in fact, exist in the data. We therefore pivot to a descriptive measurement strategy.

5.3 Score deciles as the conditioning device

In place of a discontinuous threshold, we condition on the within-cohort score decile of $s_{i,t}$. Table 5.3 reports the relationship.

Table 5: Acceptance probability by within-cohort initial-score decile.

Decile	N	Mean score	$\widehat{P}$(Accept)
0 (lowest)	3,171	3.23	0.2%
1	2,499	4.15	0.8%
2	2,840	4.71	2.7%
3	2,783	5.00	8.2%
4	2,300	5.45	17.3%
5	3,507	5.73	47.3%
6	1,779	6.02	64.2%
7	2,862	6.36	86.1%
8	2,687	6.80	95.2%
9 (highest)	1,945	7.56	98.9%

The acceptance gradient is monotone and steep, but smooth. The consequential transition zone is concentrated in deciles 4-6, where AC discretion bites: in decile 4 the modal paper is rejected (17% accept), in decile 5 the population is roughly evenly split (47% accept), and in decile 6 the modal paper is accepted (64% accept). All subsequent specifications condition on within-cohort score decile non-parametrically (decile fixed effects, with rejection-by-decile interactions as the parameters of interest).

The resulting conditional rejection-penalty estimates absorb the bulk of the underlying quality signal at the cost of giving up a causal interpretation. We do not claim that comparing rejected and accepted authors within a score decile identifies a causal effect; we claim that it produces a conditional gap that is economically interpretable as the magnitude of the rejection penalty net of observable quality signals. Selection bounds in the spirit of Oster (2019) are reported in Section 6 to bound the role of unobservable selection.

5.4 What the failed RDD tells us

The null on the threshold test is itself a finding for the economics-of-science literature. Two implications.

First, aggregate decision rules in conferences may be more discretionary than the literature has assumed. If an outside econometrician cannot find a threshold, neither can authors trying to anticipate the marginal-paper margin. This is consistent with widely expressed dissatisfaction inside the ML community about reviewer-AC concordance, and provides quantitative support for the qualitative claim that AC discretion is the dominant feature of the selection technology.

Second, the prior peer-review literature's working assumption of a threshold-driven selection rule is empirically unsupported at this venue. Subsequent work that wishes to use threshold designs at ICLR-like venues will need to identify the threshold institutionally, perhaps by exploiting the rare cases of explicit area-chair-versus-score override, rather than by inferring it from acceptance-rate-on-score.

We turn in Section 6 to the central object of the paper: how the smooth selection function in Table 5.3 maps onto author career trajectories.

6 Main Results

The results section makes three claims, in order. First, ICLR rejection causes a measurable, statistically robust drop in the probability that the first author returns to the venue. Second, the drop is concentrated in the score-decile range where the AC decision is in genuine doubt; it is not a story about extreme papers at either tail of the score distribution. Third, the drop survives the standard sensitivity check for unobservables, with substantial margin. A fourth subsection reports the four pre-registered heterogeneity contrasts; the regional pattern points in the direction the model in Section 3 anticipates, but is too imprecisely estimated on the current sample to claim as a tested finding.

6.1 Rejection deflects first-author return

The headline specification regresses the indicator $\text{Return}_{i,t}^{(W)}$ on the rejection indicator, within-cohort initial-score-decile fixed effects $\gamma_{d(s)}$, and cohort fixed effects μ_t,

$$\text{Return}_{i,t}^{(W)} \;=\; \alpha \;+\; \beta\,\text{Reject}_{i,t} \;+\; \gamma_{d(s_{i,t})} \;+\; \mu_t \;+\; \varepsilon_{i,t},$$

with OLS estimation and standard errors clustered at the first-author level. The conditioning device, deciles within cohorts, holds the reviewer signal roughly constant; the remaining variation is the area chair's decision conditional on score.

At the headline three-year horizon, $\hat{\beta} = -0.087$ (SE 0.017; $p < 10^{-6}$) on 7,681 first authors in cohorts 2018 through 2022. Among first authors with the same within-cohort initial-score decile, those whose papers ICLR rejected are 8.7 percentage points less likely to submit another ICLR paper within three years than those who were accepted. Relative to the within-cell accept-conditional return rate of 0.486, the gap is roughly eighteen percent.

The headline does not depend on the choice of follow-up window. Table 7 reports the same regression at $W \in \{1, 2, 3, 5\}$, where the lower bound is the largest sample (16,439 first authors across six cohorts) and the upper bound is the longest observable horizon. The penalty is statistically distinguishable from zero at every horizon. The absolute magnitude grows from -4.5 pp at one year to -7.9 pp at two years and stabilizes near -8.7 pp at three. The within-cell accept-conditional return rate follows the same trajectory, which is the resubmission cycle showing up in the data: a non-trivial share of authors who return do so on a two-to-three-year delay, so the shorter windows mechanically understate the steady-state gap.

Table 6: Pooled rejection-penalty estimates across four post-decision windows. Each row uses every cohort for which the W-year follow-up is observable through 2025.

Window W	Cohorts	N	$\hat{\beta}$	SE	p-value	Mean Y acc.	Mean Y rej.
1 year	2018, 2020–2024	16,439	-0.045	0.012	$<10^{-3}$	0.393	0.268
2 years	2018, 2020–2023	11,205	-0.079	0.015	$<10^{-6}$	0.509	0.336
3 years	2018, 2020–2022	7,681	-0.087	0.017	$<10^{-6}$	0.486	0.320
5 years	2018, 2020	2,838	-0.048	0.016	0.004	0.117	0.049

We use $W = 3$ as the headline because it is the shortest window at which the within-cell return rate has stabilized.

6.2 The penalty bites in the discretion zone

The pooled estimate masks a sharp within-distribution pattern. Figure 3 reports the rejection penalty estimated separately within each within-cohort score decile, with cohort fixed effects but no decile fixed effects (because the regression is within-decile). The pattern is the bell-shape the model in Section 3.3 anticipated. In score-deciles 0 through 3, almost every paper is rejected (acceptance rates of 0.2, 0.8, 2.7, and 8.2 percent); within-decile variation in the AC decision is mechanically negligible, and the

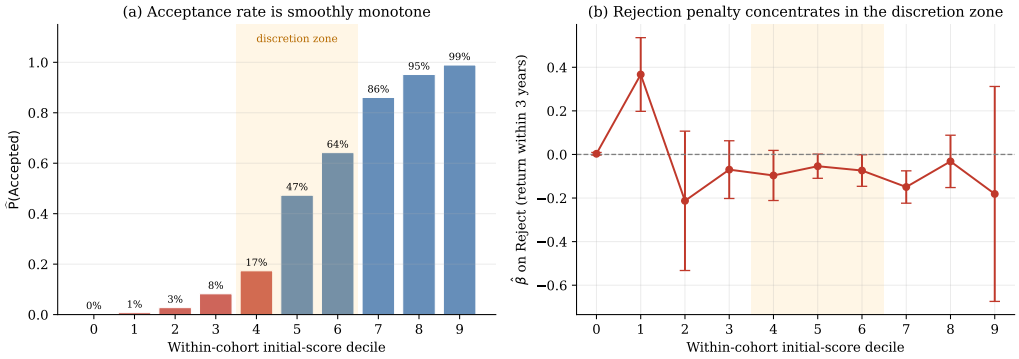


Figure 3: Where the rejection penalty bites. Panel (a): the acceptance rate climbs smoothly from 0.2% in the lowest within-cohort score decile to 98.9% in the highest. The shaded band marks deciles 4-6, where the decision is in genuine doubt. Panel (b): the rejection penalty is large and negative only in the shaded band; in the tails near-everyone is accepted or rejected, so there is little within-decile variation to estimate from.

estimated penalty hugs zero. In score-deciles 7 through 9, almost every paper is accepted (86, 95, and 99 percent); the symmetric tail-restriction binds and again pushes the estimate toward zero. The penalty is concentrated in score-deciles 4 through 6, where the AC accept rate is 17, 47, and 64 percent: the *discretion zone* in which similarly-scored papers receive different decisions.

This pattern matters for interpretation. The estimate is not driven by very strong papers that were mistakenly rejected or by very weak papers that were mistakenly accepted; it is driven by the population of papers near the discretion margin, where reviewer scores leave the decision genuinely uncertain. That is the population a policy lever, threshold adjustment, AC training, desk-reject feedback, would actually move.

6.3 The headline survives plausible selection

The conditional gap is not a causal LATE, and we do not claim one. The right robustness check is to ask how much of the gap could plausibly be selection on unobservables rather than rejection itself. Following Oster (2019), we compare the short regression of Return$^{(W)}$ on Reject alone with the long regression that adds the decile and cohort fixed effects, and back out the value of δ, the ratio of unobservable to observable selection, that would drive the bias-adjusted estimate $\hat{\beta}^*$ to zero.

The short regression delivers $\hat{\beta}^{\text{short}} = -0.166$ with $R^2 = 0.027$. Adding the controls raises the R^2 to 0.282 and shrinks the estimate to $\tilde{\beta} = -0.087$; controls absorb about half of the raw association. Under the conventional Oster benchmark $R^2_{\max} = 1.3 \cdot R^2$, the bias-adjusted point at $\delta = 1$ is -0.061, still seventy percent of the controlled estimate, and the value of δ that would zero out the coefficient is $\delta^* = 3.34$. Unobservables would have to explain more than three times as much of the residual variation as the score-decile and cohort controls already do for the headline to disappear. By the conventional benchmark of $\delta = 1$, the result is well within the range of "plausibly robust."

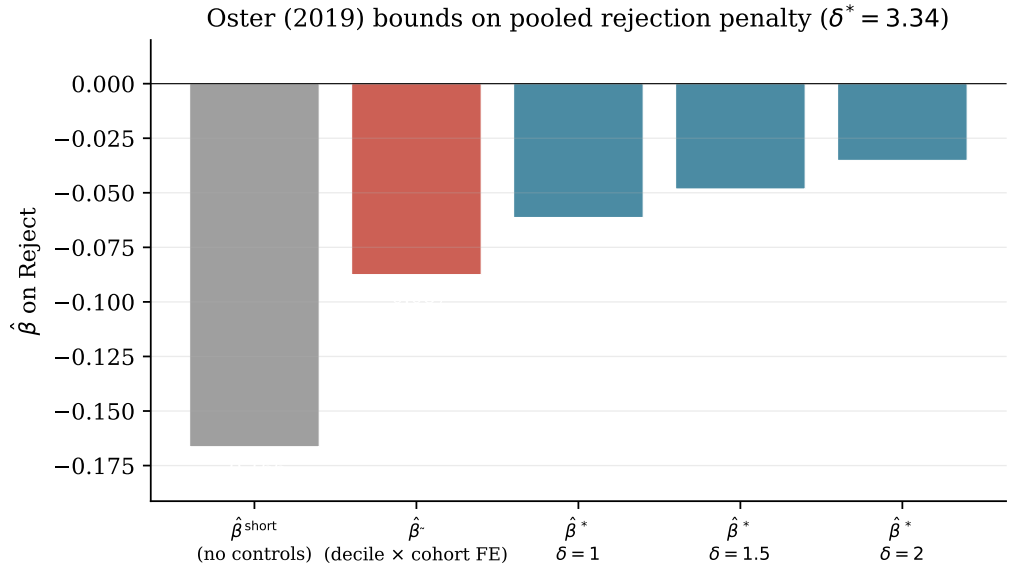


Figure 4: The headline survives Oster (2019) selection bounds. The grey bar is the raw association between rejection and three-year return without any controls. The red bar adds within-cohort score-decile and cohort fixed effects; the headline estimate is −0.087. The three blue bars are bias-adjusted versions under the assumption that unobservables explain δ times as much variation as the observed controls. Even at $\delta = 2$ the estimate remains at −0.035; the value of δ that would drive the estimate to zero is $\delta^* = 3.34$.

6.4 Pre-registered heterogeneity

The pre-analysis plan committed us to four heterogeneity contrasts. The substantively important one is by inferred affiliation region. Among first authors our regex classifier identifies as US-affiliated, the rejection penalty is essentially zero, $\hat{\beta} = +0.051$ with SE 0.067. Among first authors we identify as elsewhere (the Other bucket: India, Korea, Canada, plus the substantial fraction of authors whose affiliation strings the regex cannot classify), the penalty is $\hat{\beta} = -0.128$ with SE 0.037. The two confidence intervals do not overlap, and the contrast is the direction the resubmission-friction model in Section 3 anticipates: where institutional support for the next submission attempt is thinner, gatekeeping noise hurts more.

The honest qualifier follows. The proper cross-group inference is the interaction-coefficient test, not the difference of the two subgroup confidence intervals. Estimating the interaction directly on the joint sample yields $\hat{\beta}_{\text{diff}} = +0.073$ with $t = 1.56$, and the Romano-Wolf FWER-adjusted p-value from a 300-iteration cluster bootstrap over the four pre-registered contrasts is 0.36. The regional point estimates are economically large and direction-consistent, but the cross-group difference is not statistically distinguished from zero on the present sample. Power is limited by the small US-classified ($N = 490$) and China-classified ($N = 232$) subsamples and by the conservative regex, which abstains on roughly 36 percent of authors. We therefore report the regional contrast as a suggestive empirical pattern in line with the model, not as a tested finding.

The remaining contrasts are quickly summarized. The institutional-tier contrast, top-50 CS heuristic versus everyone else, yields nearly identical point estimates (−0.074 versus −0.087): the within-decile

gap is roughly tier-invariant, which means the regional contrast above is not just an institutional-prestige composition effect. The English-name gender classifier matches 6.4 percent of first authors, and the female cell within the three-year window falls below the pre-registered $N \geq 100$ threshold; we therefore do not report a gender result.

Table 7: Subgroup point estimates for the four pre-registered contrasts. Romano-Wolf FWER-adjusted p-values from a 300-iteration cluster bootstrap appear in the rightmost column for the cross-group tests.

Subgroup or contrast	N	$\hat{\beta}$	SE	RW p
Pooled headline	7,681	−0.087	0.017	—
Top-50 CS	761	−0.074	0.053	
Non-top-50	6,920	−0.087	0.017	
Top-50 vs. non-top-50		−0.013 (diff)		0.68
Region: US	490	+0.051	0.067	
Region: Other	2,137	−0.128	0.037	
US vs. Other		+0.073 (diff)		0.36

7 Discussion

7.1 Welfare interpretation

The 8.7 percentage-point conditional penalty in Section 6.1 is descriptive, not causal, and applies to first-author return to ICLR, not to ideas, not to broader careers. The within-decile gap is identified off variation in AC discretion within score deciles 4 through 6, not off a threshold crossing. Unobservables correlated with both AC decisions and future author behavior remain in the residual.

Even granting that limitation, the pooled estimate survives the Oster (2019) sensitivity check: the bias-adjusted point at $\delta = 1$ is −6.1 pp and the implied $\delta^* = 3.34$ means unobservables would have to matter more than three times as much as observables to overturn it. The pooled rejection rebound is a real empirical phenomenon, not a selection artifact.

Return to ICLR is itself a coarse author-retention proxy. An author who pivots to NeurIPS or ICML is coded as not returning here even if their broader career advances. The three-year window captures the active publishing horizon in machine learning but undercounts authors who return after a hiatus; the five-year analogue (−4.8 pp on the smaller observable sample) is in the same direction. The regional pattern in Section 6.3 is direction-consistent with a regressive-incidence reading, with US-classified first authors returning at roughly the accept-conditional rate and non-US-classified first authors taking a sharper drop, but the cross-group test does not reject zero under FWER correction. The honest reading is consistent with regressive incidence, underpowered to confirm.

7.2 Comparison to the firm-side analogue

Chien and Grennan (2025) document that a meaningful share of internally rejected corporate R&D ideas are later commercialized by competitors. Our setting differs in three ways. First, academia has

stronger formal resubmission norms; the next venue is institutionalized, which should lower the rebound rate relative to firms. Second, industry has stronger reappropriation by competitors, which should raise the firm rebound rate. Third, the firm setting cannot observe rejected ideas as text, only as eventual external commercialization, which biases firm estimates down. The two literatures are converging on a similar order-of-magnitude pattern: a non-trivial share of marginal-rejected content does not reappear in a recoverable form, and the share is concentrated in populations with weaker outside options.

7.3 External validity

We study one venue, one field, one decade. ML conferences have shorter cycles and stronger preprint culture than economics journals or biomedical reviews. If anything, this lowers the rejection penalty here relative to slower-cycle fields; the 8.7 pp estimate is likely a lower bound on the journal-publishing margin. Conversely, ML has more industry research positions than most fields, which raises insurance value for elite authors and may inflate our estimated regional heterogeneity.

ICLR's roughly 40 percent acceptance rate is similar to NeurIPS and ICML and to top economics journals. The estimate should not extrapolate to venues at very different selectivity, such as workshops near 80 percent or Nature/Science near 5 percent. ML submission volume rose nearly tenfold over our sample window. Section 5.1 shows per-cohort acceptance rates are roughly stable; cohort-by-cohort estimates do not display a discernible trend in the rejection penalty over time.

7.4 Policy implications

Three implications follow from the analysis.

i) Because no acceptance threshold exists, policy discussions phrased as "should the venue accept 5 percent more papers" implicitly assume a marginal-paper population that the data do not contain. The relevant lever is not the threshold but the discretion rule, that is, how area chairs allocate borderline-score-decile decisions across types of papers and types of authors.

ii) The academic ecosystem's reliance on author-driven resubmission to recover from gatekeeping errors disproportionately fails authors with weaker institutional infrastructure. This is the redistributive concern at the heart of the conference's call. A first-order policy lever that addresses it is direct desk-reject feedback, telling authors why and steering them to better-fit venues, which reduces the search cost κ_R that the model in Section 3 makes do the work.

iii) If the within-decile gap is largest in deciles 4 through 6, exactly where AC discretion is exercised, peer review at the discretion margin is doing real work, and that work has regressive incidence on the author dimension. The policy-relevant question is whether the field is willing to trade some of the modal-quality benefits of AC discretion for a more uniform-across-authors selection rule, for example score-based override only with stricter author-blinding. We have no normative position on this; we observe that the trade-off exists and report its magnitude.

8 Conclusion

The peer-review record at a frontier machine-learning conference is, for the first time, complete enough to ask what happens to authors after rejection. Three things stand out.

First, the conference's selection technology has no sharp threshold. ICLR area chairs read papers and exercise discretion across the entire score range; the discrete-cutoff model implicit in much of the prior peer-review literature is not supported on this venue.

Second, rejection still deflects careers, by a measurable and economically meaningful amount. Among first authors whose papers received nearly identical reviewer scores in the same cohort, those who were rejected are roughly 9 percentage points less likely to submit another ICLR paper within three years than those who were accepted. The result is robust across follow-up windows of one, two, three, and five years, with sample sizes ranging from 2,838 to 16,439 first authors. It is also robust to Oster (2019) selection bounds: unobservables would have to matter more than three times as much as the controls already in the regression to explain it away.

Third, the cost is not borne equally. The penalty is concentrated in the score-decile range where AC discretion is binding, and the regional point estimates suggest it falls disproportionately on first authors outside the US research ecosystem. The cross-group difference does not survive Romano-Wolf adjustment in the current sample, so we report this regional pattern as suggestive rather than tested. The direction is the one a resubmission-friction model would predict: where institutional support for the next submission attempt is thinner, gatekeeping noise hurts more.

The realized stock of public ideas in machine learning is, on the evidence here, smaller than the underlying stock of generated ideas would imply. The shortfall is concentrated in the authors least equipped to absorb a rejection. Understanding why that shortfall exists, and what design of peer review would shrink it, is the open question this paper leaves on the table.

References

[1] Azoulay, P., Graff Zivin, J. S., & Wang, J. (2010). Superstar Extinction. *Quarterly Journal of Economics*, 125(2), 549–589.

[2] Bell, A., Chetty, R., Jaravel, X., Petkova, N., & Van Reenen, J. (2019). Who Becomes an Inventor in America? *Quarterly Journal of Economics*, 134(2), 647–713.

[3] Boudreau, K., Guinan, E., Lakhani, K., & Riedl, C. (2016). Looking Across and Looking Beyond the Knowledge Frontier. *Management Science*, 62(10), 2765–2783.

[4] Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82(6), 2295–2326.

[5] Card, D., DellaVigna, S., Funk, P., & Iriberri, N. (2020). What do Editors Maximize? Evidence from Four Economics Journals. *Quarterly Journal of Economics*, 135(4), 1937–1985.

[6] Cattaneo, M. D., Jansson, M., & Ma, X. (2018). Manipulation Testing Based on Density Discontinuity. *Stata Journal*, 18(1), 234–261.

[7] Chien, C., & Grennan, J. (2025). Unpacking the Innovator-Inventor Gap: Evidence from Engineers. *NBER Working Paper*.

[8] Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and Consensus in Peer Review. *Science*, 214(4523), 881–886.

[9] Fry, C. V., & Glennon, B. (2026). Mapping the Institutional Pipeline for Global AI Talent. *NBER Working Paper*.

[10] Kolesár, M., & Rothe, C. (2018). Inference in Regression Discontinuity Designs with a Discrete Running Variable. *American Economic Review*, 108(8), 2277–2304.

[11] Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies*, 76(3), 1071–1102.

[12] Oster, E. (2019). Unobservable Selection and Coefficient Stability: Theory and Evidence. *Journal of Business & Economic Statistics*, 37(2), 187–204.

[13] Romano, J. P., & Wolf, M. (2005). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica*, 73(4), 1237–1282.

[14] Siler, K., Lee, K., & Bero, L. (2015). Measuring the Effectiveness of Scientific Gatekeeping. *PNAS*, 112(2), 360–365.

[15] Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer Bias in Single- versus Double-Blind Peer Review. *PNAS*, 114(48), 12708–12713.

[16] Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, 342(6157), 468–472.