All You Need is Q*a*L

The Crisis in Academic Publishing

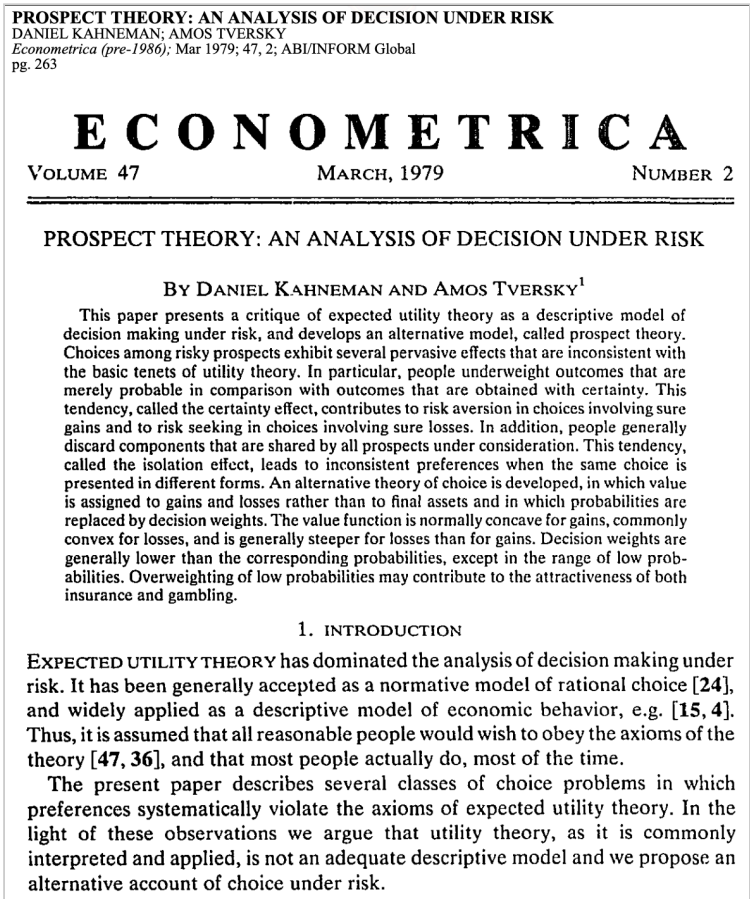
Symptoms, Diagnoses, and a Potential Solution Concept

Karl T. Ulrich *(Joint work with Gérard Cachon and Christian Terwiesch)*

July 9, 2026 · OID Department Seminar · The Wharton School

Published Articles are the Coin of the Realm

- Unit of "work" in academia
- Vary widely in embodied effort
- "Refereed" journal an important categorization
- " 'A' journal" is the most important categorization
- "Quality" matters, but assessing quality is tricky



PROSPECT THEORY: AN ANALYSIS OF DECISION UNDER RISK
DANIEL KAHNEMAN; AMOS TVERSKY
Econometrica (pre-1986); Mar 1979; 47, 2; ABI/INFORM Global
pg. 263

ECONOMETRICA

VOLUME 47 MARCH, 1979 NUMBER 2

PROSPECT THEORY: AN ANALYSIS OF DECISION UNDER RISK

BY DANIEL KAHNEMAN AND AMOS TVERSKY[1]

This paper presents a critique of expected utility theory as a descriptive model of decision making under risk, and develops an alternative model, called prospect theory. Choices among risky prospects exhibit several pervasive effects that are inconsistent with the basic tenets of utility theory. In particular, people underweight outcomes that are merely probable in comparison with outcomes that are obtained with certainty. This tendency, called the certainty effect, contributes to risk aversion in choices involving sure gains and to risk seeking in choices involving sure losses. In addition, people generally discard components that are shared by all prospects under consideration. This tendency, called the isolation effect, leads to inconsistent preferences when the same choice is presented in different forms. An alternative theory of choice is developed, in which value is assigned to gains and losses rather than to final assets and in which probabilities are replaced by decision weights. The value function is normally concave for gains, commonly convex for losses, and is generally steeper for losses than for gains. Decision weights are generally lower than the corresponding probabilities, except in the range of low probabilities. Overweighting of low probabilities may contribute to the attractiveness of both insurance and gambling.

1. INTRODUCTION

EXPECTED UTILITY THEORY has dominated the analysis of decision making under risk. It has been generally accepted as a normative model of rational choice [**24**], and widely applied as a descriptive model of economic behavior, e.g. [**15, 4**]. Thus, it is assumed that all reasonable people would wish to obey the axioms of the theory [**47, 36**], and that most people actually do, most of the time.

The present paper describes several classes of choice problems in which preferences systematically violate the axioms of expected utility theory. In the light of these observations we argue that utility theory, as it is commonly interpreted and applied, is not an adequate descriptive model and we propose an alternative account of choice under risk.

What We Mean by Quality

- Many dimensions of quality: interestingness, importance, impact, novelty, elegance, humor, clarity, rigor, popularity, methodological purity, and so on
- For simplicity, we collapse them to a single scalar Q, the eventual recognized value of the work; correctness and importance are *inputs* to Q, not separately certified axes
- Ex ante uncertainty about Q is high
- Review effort narrows that uncertainty only modestly, and possibly mostly on the dimension of correctness; **time reveals most of it**, particularly on the importance dimension

Cachon, Girotra & Netessine (2020), "Interesting, Important, and Impactful Operations Management," M&SOM 22(1)

A Metric for Q: Authority-Weighted Citations (essentially "PageRank")

- In theory could estimate Q from a wealth of independent signals, and they accrue over time, including views, citations, downloads.
- The more independent the signals, the less the estimate of Q can be gamed.
- Citations well understood and reasonably well behaved statistically; authority weighting a fairly easy add on.
- In the long run, hard to imagine a paper being important and having impact if not cited.
- (Of course there are edge cases...e.g., a paper with a widely cited methodological blunder.)

False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant

JP Simmons, LD Nelson… - Psychological …, 2011 - journals.sagepub.com

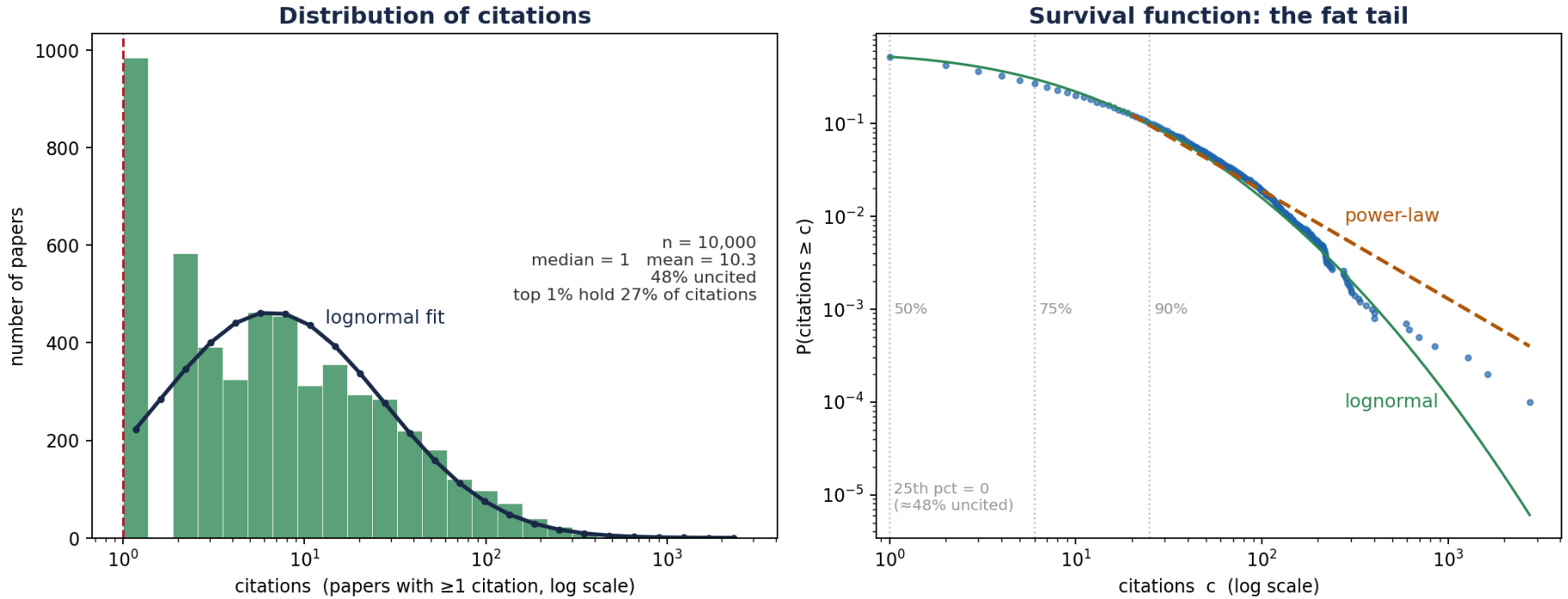
… Finally, a field known for publishing **false positives** … **false-positive** rate of 5% (ie, $p \leq .05$), current standards for disclosing details of data collection and analyses make **false positives** …

☆ Save Cite Cited by 9374 Related articles All 50 versions

The Distribution of Q Is Heavy-Tailed (and the Mode is 0)

Quality Q, proxied by citations.

Citation distribution, Management Science & Operations Research (OpenAlex, all journals · published 2015, cited through 2026 (~11-yr window) · n=10,000 sample)



Lognormal body with a power-law upper tail (cf. Radicchi et al. 2008; Brzezinski 2015; Thelwall 2016).

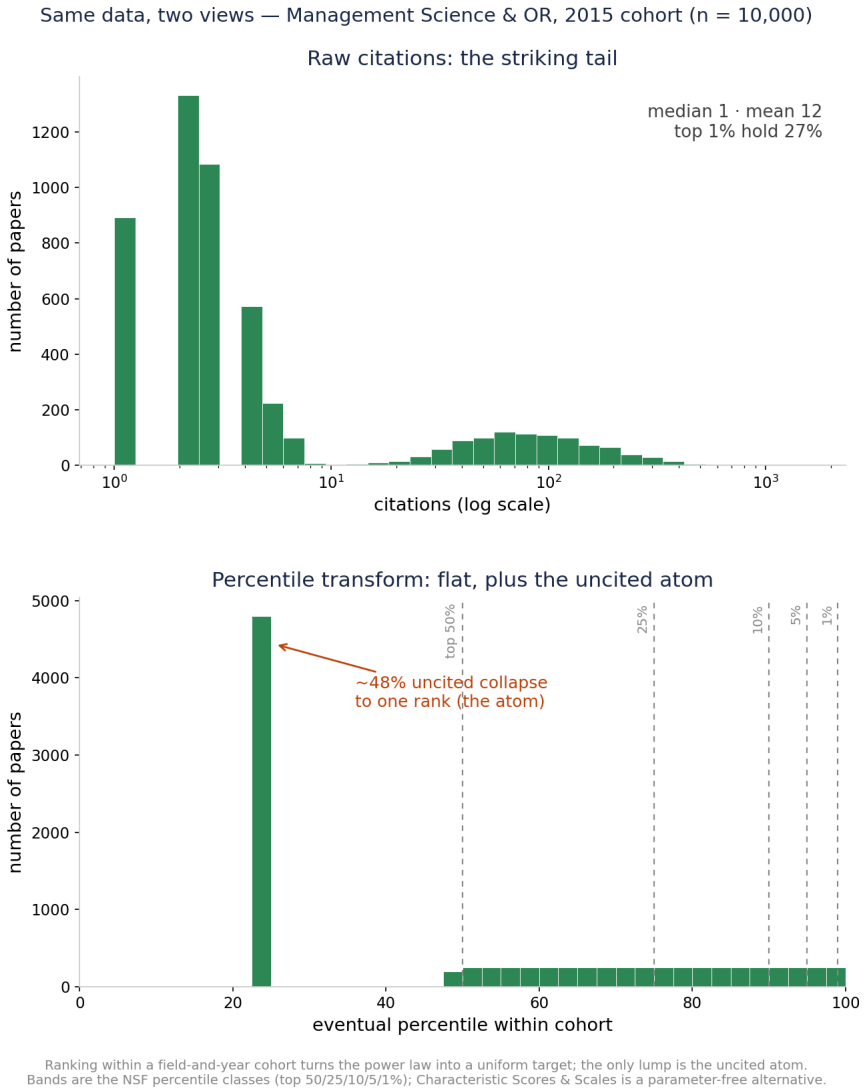
Management Science & OR, OpenAlex (all journals), published 2015, cited through 2026; n=10,000. ~48% uncited, median 1, yet the top 1% hold 27% of all citations. Psychology is essentially identical (median 0, ~53% uncited, σ≈1.5, top 1%≈26%): citation distributions are universal across fields.

Radicchi, Fortunato & Castellano (2008), *PNAS* 105(45):17268–17272; Brzezinski (2015), *Scientometrics* 103(1):213–228; Thelwall (2016), *Journal of Informetrics* 10(2):336–346.

Reporting Q as a Percentile Class

- Raw citation counts are heavy-tailed and hard to compare across fields and years
- Can report each paper in terms of **percentile within its own (sub)field and publication year**
- Report membership in classes: **top 1%, 5%, 10%, 25%, 50%**
- These are the NSF / Leiden percentile classes, a standard in research evaluation
- About half of all papers sit at or near zero citations: a mass at the bottom

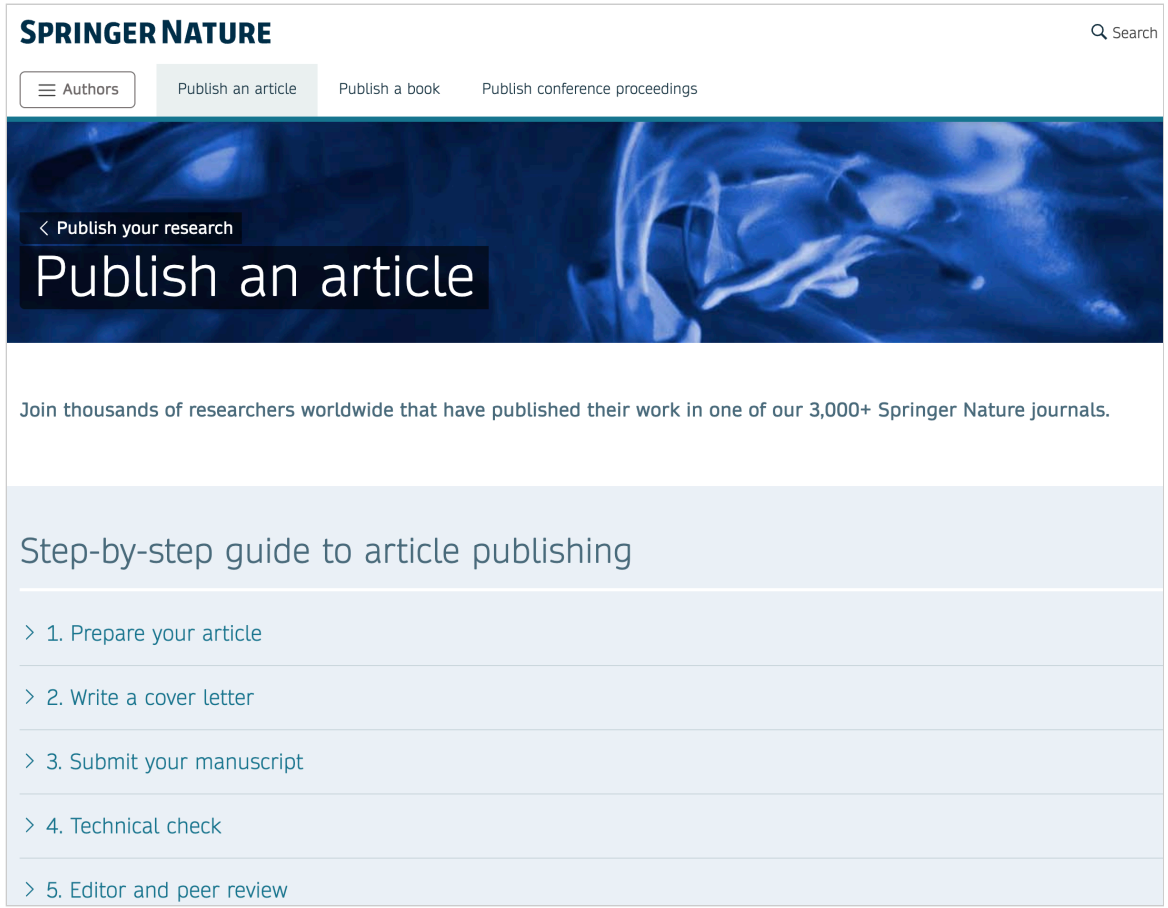
A percentile is intuitive, field-fair, and robust to the tail.



Percentile classes: National Science Board, *Science & Engineering Indicators* (NSB-2025-7); CWTS Leiden Ranking. Parameter-free alternative: Glänzel & Schubert (1988), "Characteristic Scores and Scales in Assessing Citation Impact," *Journal of Information Science* 14(2):123–127.

The Societal Functions of Publishing

- **Administration**: time stamp, copy edit, typeset, archive, distribute
- **Development**: peer feedback and guidance for revision
- **Promotion**: communicating existence to the relevant audience
- **Certification of quality**: branding the work as a signal of quality



Historically, Journals Performed All of these Functions

- Most functions (except development) were completed nearly **simultaneously**, around the time of "publication"
- Today the functions are increasingly **unbundled** (SSRN, arXiv, personal sites, social media, DIY publication-quality document prep, LLM copy editing)
- Administration is now largely irrelevant (PDF downloads, SSRN, open access)

The functions of scholarly communication were canonically identified as registration, certification, awareness, and archiving by Roosendaal & Geurts (1997), "Forces and Functions in Scientific Communication"; the case for *unbundling* (decoupling) them into independent services was made by Priem & Hemminger (2012), *Frontiers in Computational Neuroscience* 6:19.

The one thing Journals purport to still do (certification of quality) can't really be done, for two reasons

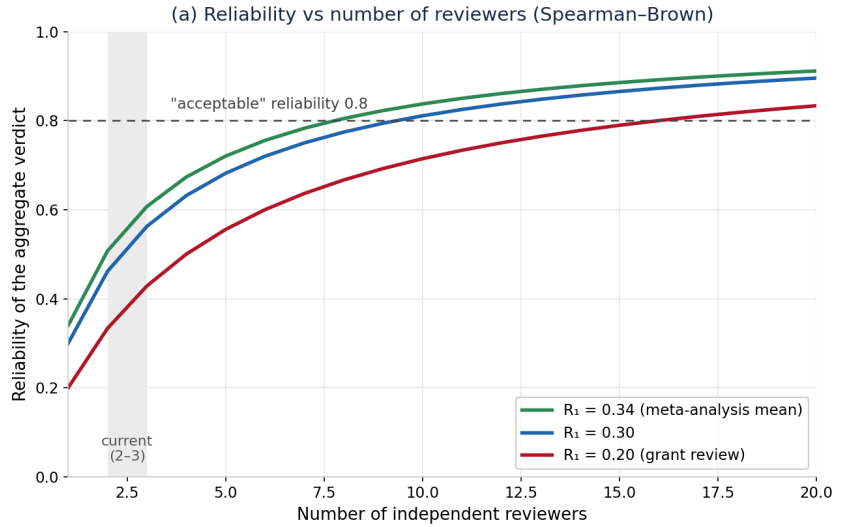
1. The instrument is noisy

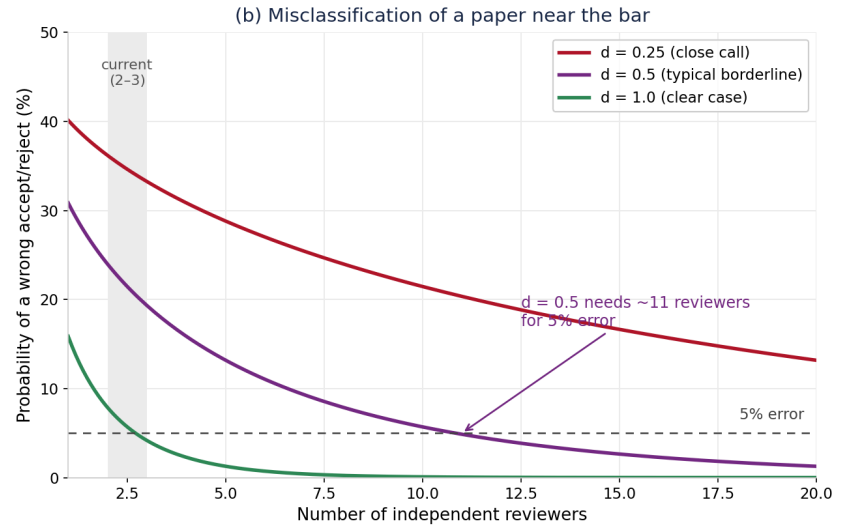
- Inter-rater reliability is low: 2–3 reviewers give a verdict little better than a guess
- Reliability rises only slowly with more judges; you would need >10 to dependably classify a marginal paper
- *Fixable in principle*: add judges (Spearman–Brown)

2. The target is mostly exogenous

- Even a perfect read on a paper's knowable quality predicts little of its eventual Q
- Most of what determines eventual Q (timing, attention, who cites whom) is realized **after** the decision
- *Not fixable* by any number of judges

Reason 1 limits any small panel. Reason 2 limits any ex ante process, however many reviewers it uses.





A Simple Model, with Evidence for Each Parameter

Eventual quality decomposes into a knowable component plus an exogenous shock; each reviewer sees the knowable component through noise:

$$Q = q_0 + u \cdot s_i = q_0 + e_i$$

Parameter	Value	Evidence
Reviewer noise (reliability of one review)	ICC ≈ 0.34	Bornmann, Mutz & Daniel (2010), *PLoS ONE* 5(12):e14331 (48 studies, 19,443 manuscripts); Cicchetti (1991), *Behavioral & Brain Sciences* 14(1):119–186; NeurIPS revisit, ~50% of scoring subjective — Cortes & Lawrence (2021), arXiv:2109.09774
Validity ceiling (knowable quality → eventual Q)	R^2 ≈ 0.06	Success in cumulative-advantage markets is intrinsically unpredictable — Salganik, Dodds & Watts (2006), *Science* 311:854–856; review scores do not predict citations of accepted papers — Cortes & Lawrence (2021); impact often realized years later ("sleeping beauties") — Ke, Ferrara, Radicchi & Flammini (2015), *PNAS* 112(24):7426–7431

q_0 = the knowable (ex ante) component of quality; u = the exogenous shock realized after the decision, so eventual quality $Q = q_0 + u$; e_i = reviewer error. A review recovers at best q_0; even perfectly, it explains only about 6% of Q. Outcomes are binned into the NSF percentile classes; the validity ceiling is sourced from the publishing literature, not from our own idea-evaluation work.

The Result: More Reviewers Barely Help

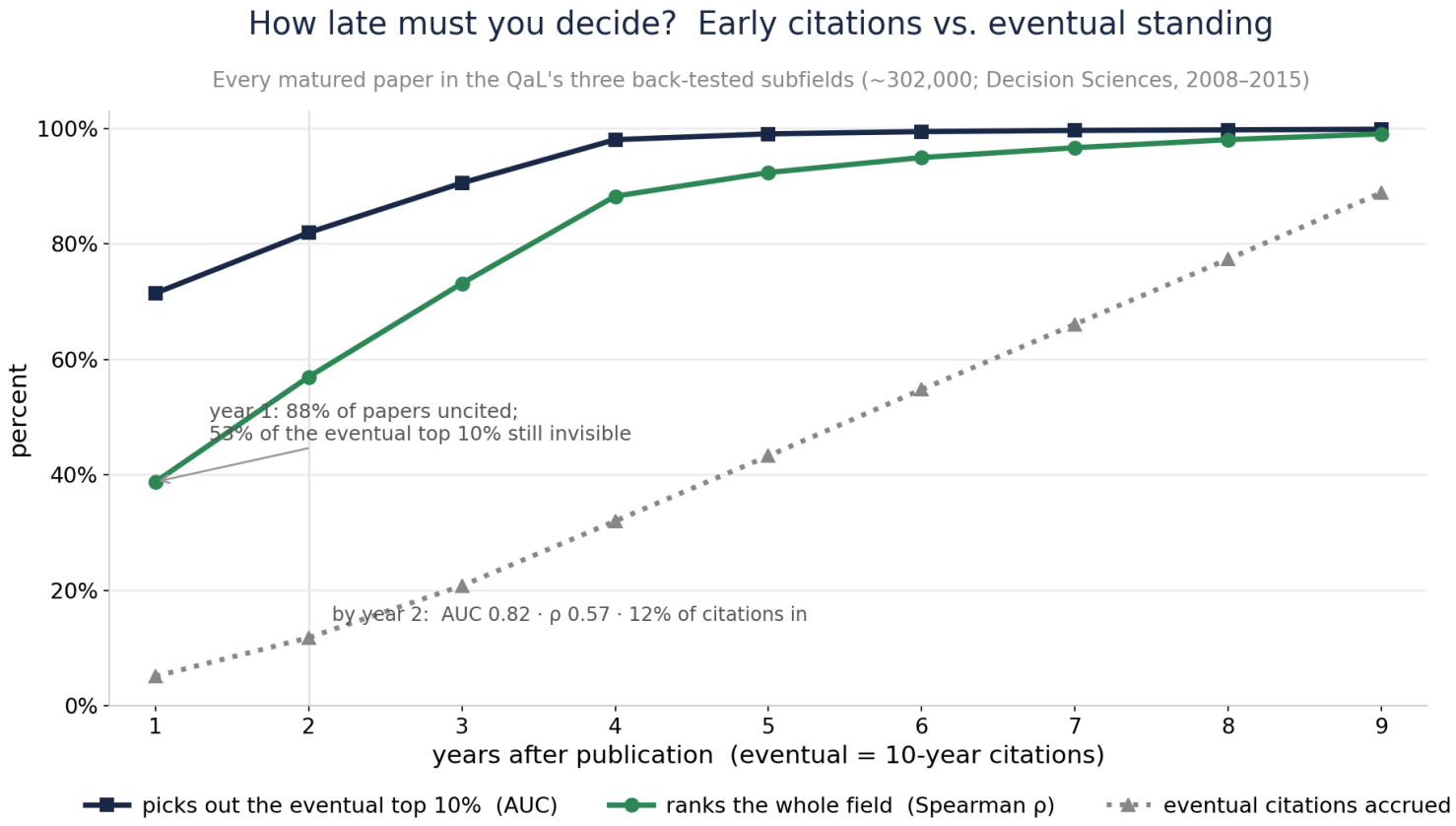
Probability of correctly identifying	By chance	3 reviewers	10 reviewers	Perfect read of q_0
a top 1% paper	1%	3%	4%	**4%**
a top 10% paper	10%	17%	18%	**19%**
a top 25% paper	25%	33%	35%	**36%**
a top 50% paper	50%	56%	57%	**58%**

Probability a paper that truly belongs in a tier is placed there, under the model above (ICC 0.34; validity ceiling R^2 0.06). The last column is a perfect (noise-free) read of q_0 — the knowable ex-ante quality defined on the previous slide ($Q = q_0 + u$) — **not** foreknowledge of the outcome; even then a top-1% paper is tagged only ~4%, because ~94% of impact is exogenous. Going from 3 to 10 reviewers barely moves the numbers: the validity ceiling, not reviewer count, is the binding constraint.

Early Citations Predict Eventual Rank

- Sample of 302k papers behind the QaL's calibration (three Decision-Sciences subfields, 2008–2015; eventual = the 10-year percentile)
- Top 10% in year N as a predictor of eventual top 10%: **AUC 0.82 by year 2**, **0.98 by year 4** Even 2-year citations likely much better than 3 reviewers ex ante.
- The atom at zero (88% uncited at year 1) flatters a plain rank correlation (Spearman ρ 0.39) and floors the tie-corrected one (Kendall τ-b 0.20); AUC (area under curve for binary classifier) ignores the uncited mass and is best calibration metric.
- Number of citations matures slowly — only 12% of eventual citations are in by year 2, 43% by year 5

Real signal early, full picture late.



The System has Always been Infuriating; It May Be Getting Worse

- Supply is increasing dramatically
- LLM use in writing
- Co-authorship inflation (no obvious net effect on the number of papers)
- Submissions to *Organization Science* up **42%** since ChatGPT (a COVID bump was 20%), almost entirely AI-generated text
- Of the most AI-heavy manuscripts (70%+), ~70% are desk-rejected, vs 44% for low-AI
- Reviewers increasingly rely on LLMs for referee reports (a "shadow AI review system")
- Does not necessarily make the system worse, but we can probably use AI more deliberately
- Anecdotally, time to first review is increasing, but variance across journals is very high

Pierce, Gartenberg, Murray & Hasan, "More versus Better," Organization Science (2026)

What are the stakeholder needs for an academic publishing system?

(The Clean-Sheet Perspective)

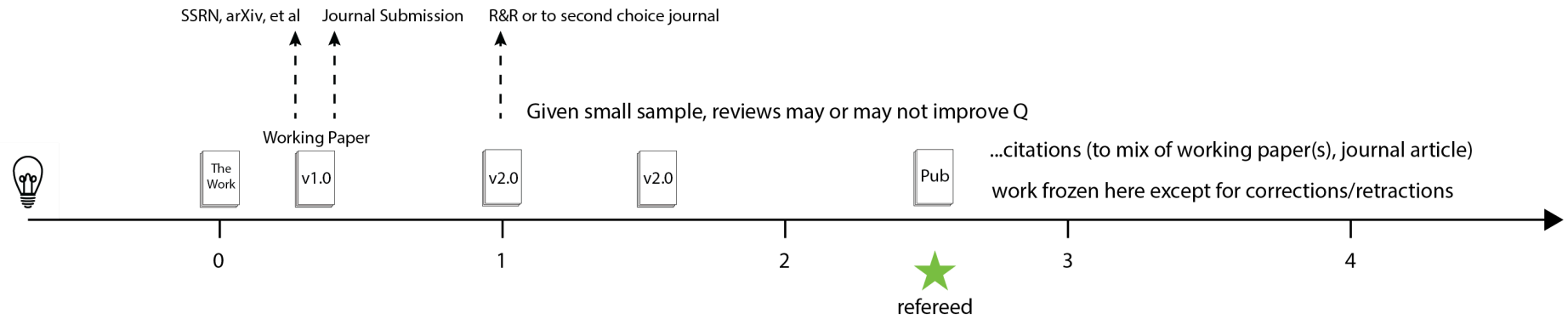
- Credibly certify correctness, avoiding false negatives
- Provide reliable, unbiased, hard-to-game evidence of impact
- Minimize the time from submission to "reviewed publication" CV status
- Allow papers to be "reviewed/published" while minimizing unproductive author effort
- Engage the community in discovery, discussion, and constructive feedback
- Embrace revision over time with a transparent history and serving the latest version
- Exhibit grace towards low-citation papers and authors (the vast majority of the field)
- Enable easy discovery, search, and retrieval
- Operate in an economically sustainable way

A cursory tour of academic forums (Hacker News, r/AskAcademia, r/professors) turns up every one of these, in far more colorful form. On what the community wants: Mulligan, Hall & Raphael (2013), *JASIST* 64(1):132–161; Ithaka S+R *US Faculty Survey* (2021).

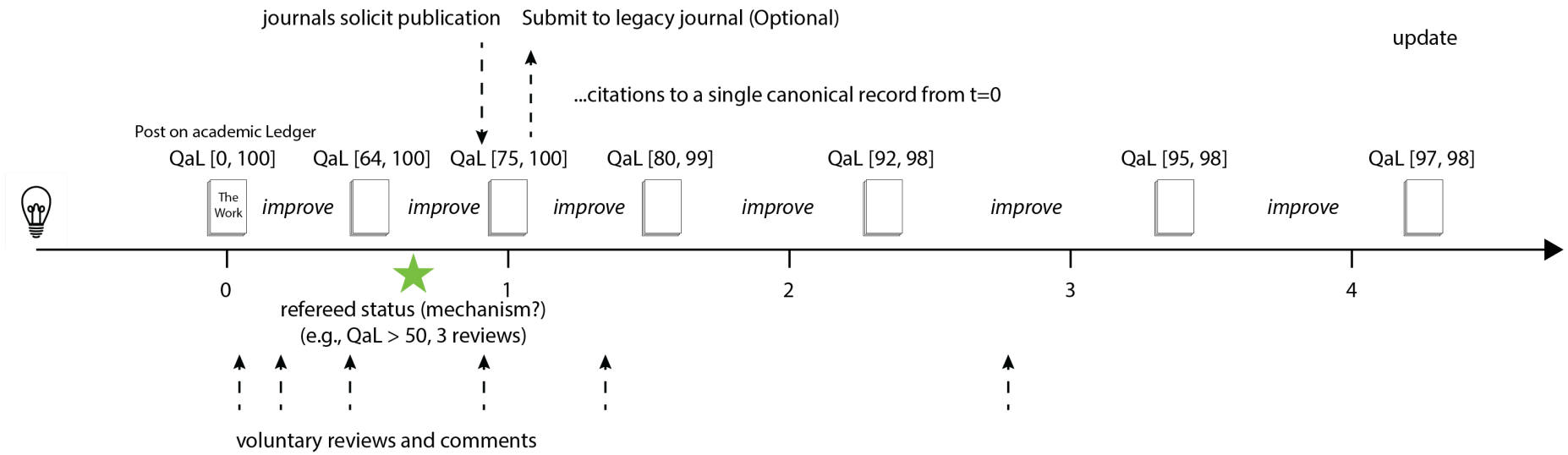
An Alternative Approach: The Academic Ledger

- A neutral, non-profit *system of record* applies a light integrity screen **immediately** and enters the work in a ledger, without judging quality
- Then evidence of quality **accumulates over time**, and the branding of quality occurs with some delay, when confidence in the estimate of Q is sufficiently high
- Legacy journals co-exist and may remain important for branding and curation around topics, but the Ledger is the source of truth about quality.
- Alternatives to conventional journals may emerge, built on top of the Ledger to deliver other benefits (e.g., engagement with a community, improvement of the work, focused curation around topics)

Existing Timeline - Legacy Journals



Possible Timeline - academic Ledger



Beta Build of academic Ledger (with QaL)

- Built using OpenAlex as the source data.
- Estimates QaL for virtually any paper in any field (N ~ 500M)
- Basic approach is
    - Construct a synthetic subfield for any paper (using hierarchy of methods depending on paper age)
    - Locate the paper in the synthetic subfield in percentile rank relative to subfield-year cohort.
    - Percentile rank in each subfield is exact...an observation. Combined by weighting subfields.
    - Provide a 90 percent confidence interval of eventual QaL (calibrated by observing papers in subfields)

Ideas are Dimes a Dozen: Large Language Models for Idea Generation in Innovation

Karan Girotra, Lennart Meincke, Christian Terwiesch, Karl T. Ulrich

SSRN Electronic Journal · Exp & Cog Psych 21% · +7 more · Posted **2023** · 3 years on the record

QUALITY ON THE ACADEMIC LEDGER

QaL

97.3 **99.8** 100

eventual percentile · point estimate with 90% interval · *illustrative*

Reference class: synthetic field · official headline recency-weighted blend of its community's cohorts · 74% back-tested · field pct 99.89

probability of landing in each NSF percentile class (top 50 / 25 / 10 / 5 / 1%)

Decided late: the interval is wide early and narrows as evidence accrues.

SYNTHETIC FIELD · WEIGHTED BY RECENT REFERENCES

Experimental and Cognitive Psychology (21%) · Artificial Intelligence (17%) · Sociology and Political Science (11%) · Health Informatics (10%) · Computer Science Applications (5%) · Economics and Econometrics (5%) · Management Information Systems (4%) · Cognitive Neuroscience (4%)

The official reference class is this recency-weighted blend of the communities the paper actually draws on (QaL_spec §5), not OpenAlex's single label ("Computer Science"). Divergence is the signal.

EVIDENCE BEHIND THE ESTIMATE — pulled from OpenAlex

168 Citations (total)	**99.51%** Observed percentile (synthetic field)	**None** Retractions / corrections
3 yr Age since posting	**Computer Science** Detected field (OpenAlex)	

Designing Q*a*L: Composition and Choices

How it is composed

- A blend of **robust, article-level, subfield-and-vintage-normalized** signals, never a raw count
- Value is the paper's **percentile within field and publication year** (NSF / Leiden classes)
- Synthetic subfield constructed with **co-citation-neighborhood normalization** (Relative Citation Ratio) and subfield-normalized scores (Leiden MNCS)
- **Authority-weighted** citations (PageRank-style), harder to game than raw counts (not yet implemented)
- Correctness signals fold in over time: replications, corrections, retractions (not yet implemented)
- Reported as a point estimate, a 90% interval, and bucket probabilities, all of which evolve with additional information

Design questions

- **Transparent or opaque?** Transparent method and inputs, auditable and reproducible; opaque metrics breed distrust. Gaming-resistance comes from *design*, not secrecy: many independent signals, authority-weighting, a percentile basis, and deciding late
- **Field-normalized how?** Within field × vintage, plus article-level co-citation normalization; the reference class is always stated
- **Proven, robust tools to invoke:** percentile classes, field-normalized scores (MNCS), the Relative Citation Ratio, Characteristic Scores and Scales — not the article-level Journal Impact Factor

Responsible-metrics principles: Hicks, Wouters, Waltman, de Rijcke & Rafols (2015), "The Leiden Manifesto for research metrics," *Nature* 520:429–431; DORA (2013). Article-level normalization: Hutchins, Yuan, Anderson & Santangelo (2016), "Relative Citation Ratio," *PLoS Biology* 14(9):e1002541; CWTS Leiden Ranking (MNCS); Glänzel & Schubert (1988).

Serving Democracy: Evidence of Voting Resource Disparity in Florida

Gérard P. Cachon, Dawson Kaaua

Management Science · PoliSci & IR 50% · +7 more · Posted **2022** · 4 years on the record

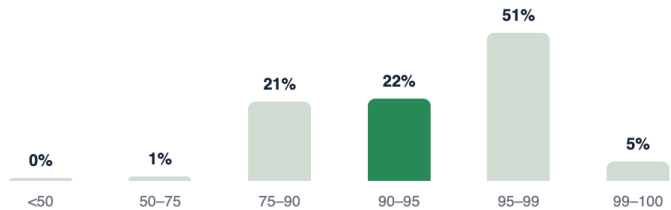
QUALITY ON THE ACADEMIC LEDGER

QaL

89 **94** 98

eventual percentile · 90% interval · *preliminary — this community isn't back-tested yet*

Reference class: synthetic field · official headline recency-weighted blend of its community's cohorts · preliminary · 34% back-tested · field pct 97.79



probability of landing in each NSF percentile class (top 50 / 25 / 10 / 5 / 1%)

Decided late: the interval is wide early and narrows as evidence accrues.

SYNTHETIC FIELD · WEIGHTED BY RECENT REFERENCES

Political Science and International Relations (50%) · Management Information Systems (15%) · Artificial Intelligence (10%) · Sociology and Political Science (8%) · Public Administration (6%) · Transportation (5%) · Economics and Econometrics (4%) · +1 more (1%)

The official reference class is this recency-weighted blend of the communities the paper actually draws on (QaL_spec §5), not OpenAlex's single label ("Social Sciences"). Divergence is the signal.

EVIDENCE BEHIND THE ESTIMATE — pulled from OpenAlex

13 Citations (total)	**94.6%** Observed percentile (synthetic field)	**None** Retractions / corrections
4 yr Age since posting	**Social Sciences** Detected field (OpenAlex)	

Computing a QaL: A Worked Example

Subfield in the synthetic field	Weight	Pctile	Contribution
Political Science & International Relations	37.7%	97.8	36.9
Management Information Systems	15.3%	91.0	13.9
Economics & Econometrics	10.9%	93.7	10.2
Artificial Intelligence	8.2%	86.7	7.1
Sociology & Political Science	8.0%	95.2	7.6
Strategy & Management	7.3%	90.8	6.6
Public Administration	5.2%	93.0	4.8
Transportation	4.4%	88.6	3.9
Marketing	3.0%	86.2	2.6
Total	**100%**		**93.7**

Contribution = weight × percentile; the paper is ranked within each subfield's 2022 cohort. The subfields disagree (86th–98th) — averaging them is what pulls the blended percentile down to 93.7.

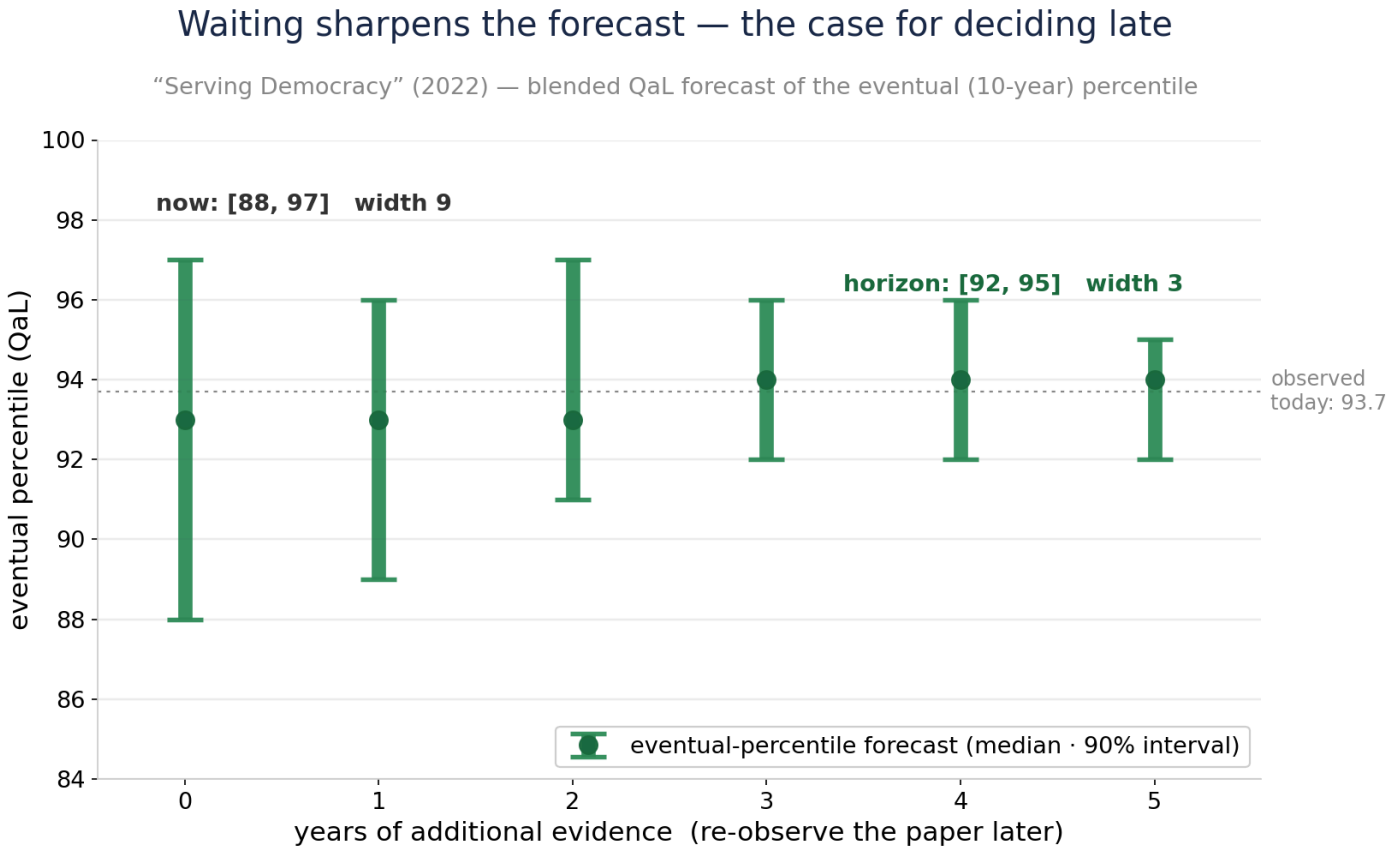
Serving Democracy — voting-resource disparity in Florida, Cachon and Kaaua (2022); 13 citations.
OpenAlex tags it **one** subfield (*Political Science & International Relations*). QaL instead ranks it in the **synthetic field** — the 9 subfields its references and co-citations actually span.
Observed = Σ (weight × percentile) = **93.7**
Extrapolate to eventual (its percentile at the 10-year horizon) — it's only 4 years old, so each calibrated subfield maps observed → eventual, blended over the 95% of weight that's calibrated:

QaL 93 · 90% CI [88, 97]

a likely top-10% paper — the eventual standing is a genuine **forecast**, with an interval that **narrows as the paper matures**, not just today's citation count.

Forecasting from Now to the Horizon

- A young paper's cohort **isn't mature** — citations are still accruing, so today's rank isn't its final one
- Calibration — learned from communities we have watched to **maturity** — maps *observed percentile + age* → a **distribution** over the eventual percentile
- The **90% interval narrows** as the paper ages toward the 10-year horizon, [88, 97] → [92, 95] — more evidence, not a different answer
- That shrinking uncertainty is the case for **deciding late**: the eventual standing is a **forecast** you can sharpen by waiting, so commit when the evidence is in — not at submission.
- Still, the estimate is pretty good with a few citations.



How Might we Provide "Refereed" Status with Minimal Friction

- "Incidental" voluntary reviews. (e.g., signed comments from readers who are already interested organically in the work: (a) what did I most like about this paper?, (b) how could the author make this paper better?)
- Existing journals could solicit articles from work posted on the Ledger.
- Emergent virtual journals with volunteer editorial boards for a very specific topic. (I might actually want to be a reviewer for a Journal of Innovation Processes.)
- Automated LLM feedback using best current practices, posted with information on prompt and model.

Articles in Refereed Journals

[13] Karl T. Ulrich, "Digital Darwinism: Steering the Evolution of Artificial Life in Socio-Technical Systems," *AI and Ethics* (2026) 6:268. https://doi.org/10.1007/s43681-026-01057-8

[14] Karl T. Ulrich, "Re-Consider the Lobster: Animal Lives in Protein Supply Chains," *Sustainability*, 17, 2025, p. 7034. https://doi.org/10.3390/su17157034

[15] Karl T. Ulrich, "Minimum Spatial Housing Requirements for Human Flourishing," *Buildings*, 15, 2025, p. 2623. https://doi.org/10.3390/buildings15152623

Working Papers including Papers under Review or Revision

[48] Ulrich, Karl T. "The Satoshi Overhang: Why the Bear Case is Bounded." *under review*. https://doi.org/10.48550/arXiv.2604.27694

[49] Ulrich, Karl T. "Why is Skiing Fun? A Compound Superstimulus Theory of Risky Effortful Movement," *under review*.

[50] Girotra, Karan and Meincke, Lennart and Terwiesch, Christian and Ulrich, Karl T., Ideas are Dimes a Dozen: Large Language Models for Idea Generation in Innovation. Report – Mack Center for Technological Innovation. University of Pennsylvania (July 10, 2023). Available at SSRN: https://ssrn.com/abstract=4526071

[51] Kornish, Laura J. and Karl T. Ulrich, "Assessing the Quality of Selection Processes," under revision for resubmission, March 2017.

[52] Wooten, J. and K. Ulrich, "The Impact of Visibility in Innovation Tournaments: Evidence from Field Experiments," under revision for resubmission, March 2017.

[53] Terwiesch, Christian and Karl T. Ulrich, "Will Video Kill the Classroom Star? The Threat and Opportunity of Massively Open On-Line Courses for Full-Time MBA Programs," Report – Mack Center for Technological Innovation. University of Pennsylvania, July 2014.

[54] Girotra, K. and K. Ulrich, "Empirical Evidence for Domain Name Performance," Working Paper, The Wharton School, May 2011.

[55] Ulrich, Karl T., "The Environmental Paradox of Bicycling," Working Paper, The Wharton School, July 2006.

The Legacy Journal Perspective

- The Ledger is consistent with the policies of most academic journals in social science
- Journals could encourage authors to **first publish on the Ledger**, reducing their volume of low-quality submissions
- The alternative to a hard reject could become "develop the work on the Ledger"
- Editors could proactively scan the Ledger for work to invite, evaluating fit and quality based on evidence

Utopia

- Administration is smooth and fast: register, screen integrity, archive, timestamp, and disseminate within days
- True signals of quality emerge **smoothly and continuously** as the field reads, uses, cites, and endorses; those signals are used for critical decisions in universities
- Authors can **respond, adapt, and improve** the work, with the revision history on the record
- Incentive is to create interesting and impactful work not to fit the norms of a community
- Journals become nice-to-have *curation*; the revealed evidence is what matters for reputation and for evaluating scholarship
- A more predictable, less capricious, less arbitrary world for emerging scholars.
- Senior scholars with no patience for journals may opt out of the legacy system and still "publish."

Some Key Questions

- Does the system require radical change, or could the incumbents address the pain points?
- Could an institution host the Ledger (e.g., "academic Ledger is hosted by the Wharton School of the University of Pennsylvania")?
- What is the unit of work? (papers, books, videos, other) Should we care?
- What resources are required to operate the Ledger at scale?
- What business models might work to make the Ledger sustainable?

OTHER STUFF

The Initial Screen Should be a Low Bar

- **Entry in the Ledger does not judge quality.** It verifies identity and screens out fraud and the obviously wrong, nothing more
- **Identity is the primary filter:** ORCID authentication raises the cost of paper-mill, sock-puppet, and fabricated-author submissions
- The automated and LLM screen targets a deliberately narrow set: fabrication, manipulated figures, paper-mill and "tortured phrase" text, fabricated or dead citations, gross statistical inconsistencies, and claims with no support
- Much of this is handled by **validated rule-based tools** (image-duplication, statcheck, tortured-phrase and paper-mill detectors), with LLMs for triage. This is a low bar the evidence can support, unlike judging quality, which it does not attempt
- Expected outcome: outright rejection of perhaps **1–5%** of ORCID-authenticated submissions
- **Prefer conditional registration:** admit borderline work with an attached, transparent summary of the LLM-generated concerns, and leave judgment to readers and to the later tiers

LLMs cannot yet reliably certify correctness (Son et al. 2025, SPOT; Xi et al. 2025, FLAWS; Dycke & Gurevych 2025) and are gameable (Gibney 2025, *Nature* 643:887). The screen is therefore scoped to fraud and obvious error, where rule-based tools (statcheck; image-duplication; tortured-phrase and SCIgen detectors) are already effective.

How Is This Not Just SSRN?

- SSRN is a for-profit organization, owned by Elsevier (a journal publisher), with conflicted incentives
- SSRN is a posting service, not a system of record
- SSRN, despite offering its own branded metric (PlumX), deliberately avoids an honest estimate of Q
- It confers no credential that "counts," its download metrics have been gameable, and are reported only as a running cumulative total
- The Ledger adds what SSRN lacks: neutral non-profit governance, an identity and integrity layer, signed review, "refereed" status, and an evidence-based estimate of Q in percentile terms

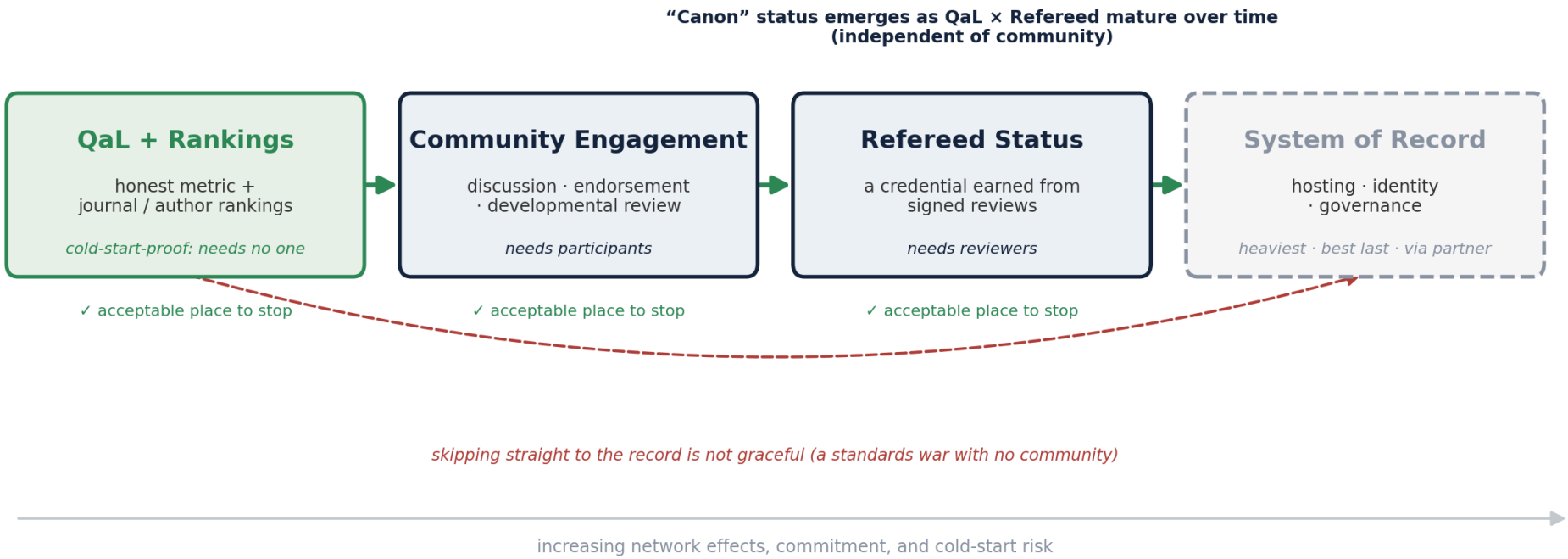
Could an Incumbent Implement the Ledger?

- An incumbent doing this well would be a **win, not a loss**. The aim is the system of record with a credible estimate of Q, by whatever means best realizes the goals.
- What a record requires: **neutral, non-commercial governance**, permanence, coverage of all fields, an identity and integrity layer.
- Barriers for anyone: the cold start (authors and endorsers together); recognition still locked to journal names and the impact factor; cross-field scope; durable funding; each incumbent's own inertia.

Player	Strengths as a record	Weaknesses / risk
arXiv (independent non-profit, from Jul 2026)	Neutral, trusted, built to last; already spans physics to economics	Identity tied to physics/math/CS; no endorsement, curation, or metrics layer
SSRN (for-profit; Elsevier, 2016)	Deep social-science adoption and author base	Owned by a journal publisher; conflicted; wrong steward for a neutral record
bioRxiv / medRxiv (openRxiv, non-profit, 2025)	Strong governance, fast growth, well funded (CZI)	Life and health sciences only
eLife (non-profit journal)	Already layers signed review on preprints (our Refereed idea)	A journal, life-science scope; lost its impact factor when Clarivate balked
Octopus (UKRI-funded non-profit)	Explicitly a "new primary research record"	Radical unit-of-research model; thin adoption; UK-centric
Big publishers (Elsevier, Springer Nature, Clarivate)	Scale, money, distribution	Most conflicted; their economics oppose decoupling

Adjacent pieces exist but none certify or curate: OpenAlex, Crossref, ORCID (non-profit data and identity); Google Scholar, Semantic Scholar (discovery); ResearchGate (for-profit network).

Cold Start and Graceful Evolution



Four non-overlapping feature sets, drawn as states. **QaL + Rankings** is cold-start-proof and bootstraps the rest; **Community Engagement** supplies the reviews that earn **Refereed Status**; the **System of Record** is heaviest and comes last. Each of the first three is an acceptable place to stop, so the system stays additive to SSRN, arXiv, and journals and degrades gracefully. Canon is not a separate set: it emerges as QaL and Refereed mature.

The Journal System Could Possibly Be Fixed

- We have faced long lead times and Reviewer 2 essentially forever. Not a new problem.
- Tightly run journals with committed editors still seem to function reasonably well
- "AI will save us" is the optimistic branch, but the evidence is mixed and the prospects are genuinely uncertain
- LLMs are useful *assistants*: on realistic tests they catch only a minority of real or inserted errors, miss broken logic, can be gamed by hidden prompts, and skew lenient. They may get much better.
- They are strongest at triaging and improving reviews, weakest at identifying the very best work and at grounded technical verification
- A diverse **ensemble** of LLMs (the "wisdom of crowds" move) may raise the floor cheaply, but the evidence does not yet support replacing human judgment
- Still, journals could use LLMs for an initial filter to push net submission rates toward pre-LLM levels

LLMs catch only roughly 20–40% of real or inserted errors and can miss broken logic entirely: Son et al. (2025, SPOT, arXiv:2505.11855); Xi et al. (2025, FLAWS, arXiv:2511.21843); Dycke & Gurevych (2025, arXiv:2508.21422). Useful as assistance: Liang et al. (2024), *NEJM AI* 1(8). Gameable by hidden prompts: Gibney (2025), *Nature* 643:887–888. On the ensemble idea: Meincke, Terwiesch & Huchzermeier (working paper).

Three Sequential Categories of Papers: Certified, Refereed, Canon

- **Certified.** Authenticated by ORCID (the primary filter) and screened by automated tools and LLMs for fraud and obvious defects, not correctness, then archived, timestamped, and public within days. A floor against fraud and the obviously wrong, not a claim about quality. Fast and near-automatic; perhaps 1–5% are rejected.
- **Refereed.** Has earned named, signed reviews or endorsements from identified scholars above a published threshold. The tier that honestly merits the classification "peer reviewed," but in the open and on the record. Perhaps the standard is 90% confidence that the paper will eventually reach the 75th percentile of submissions.
- **Canon.** The curated, time-earned best, conferred from realized impact (use, citation, replication, discussion) and signed endorsements, by field panels under published criteria. Quality judged after the fact from evidence, not predicted. Perhaps approximately the 95th percentile of papers. Because it rests on evidence rather than a noisy verdict from two or three reviewers, it should in theory be more valuable than journal acceptance.
- **Published.** Optional nice-to-have curation via publication in a journal. Publication could, but need not, follow from the evidence.

One record: a CV line begins as *Certified* and is upgraded in place to *Refereed*, then *Canon*, as review and impact accrue. A credential that only appreciates.

The Same Paper, Several Ways

`[Certified]` Cachon, G., C. Terwiesch, and K. Ulrich. 2026. The Crisis in Academic Publishing. *Academic Ledger* [0, 100]. https://doi.org/10.59312/aldg.2026.0427

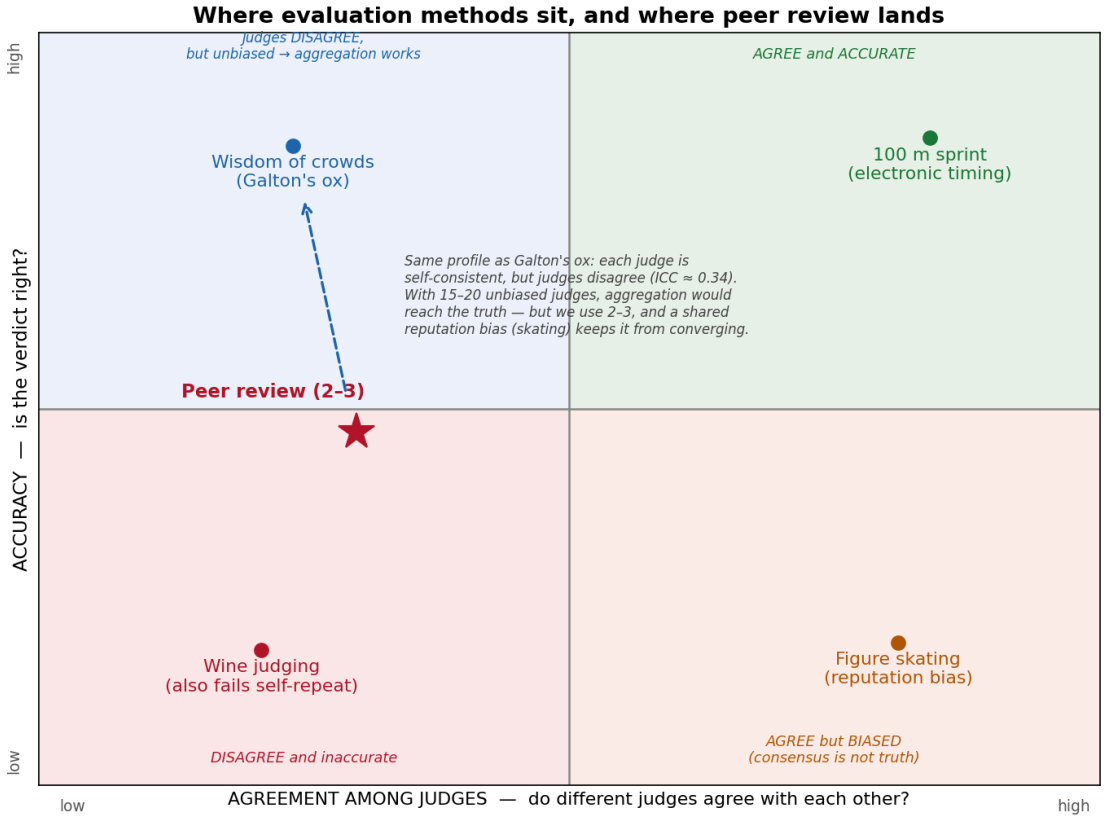
`[Refereed]` Cachon, G., C. Terwiesch, and K. Ulrich. 2026. The Crisis in Academic Publishing. *Academic Ledger* [75, 100]. https://doi.org/10.59312/aldg.2026.0427

`[Canon]` Cachon, G., C. Terwiesch, and K. Ulrich. 2026. The Crisis in Academic Publishing. *Academic Ledger* [95, 100]. https://doi.org/10.59312/aldg.2026.0427

`[Canon]` Ⓟ Cachon, G., C. Terwiesch, and K. Ulrich. 2026. The Crisis in Academic Publishing. *Academic Ledger*. (also appears in Quantitative Science Studies 7(2):512–540, 2026). https://doi.org/10.59312/aldg.2026.0427

One record, one DOI, one act of authorship: the line is upgraded in place; the circled P flags the journal-published version, with its citation appended.

Wine, the 100m Dash, Galton's Ox, and Figure Skating



Modest Success ("SSRN/arXiv+")

A realistic near-term target: a system of record, plus…

- An automated fraud and integrity screen (a low bar, not full correctness) and certification
- Community engagement, discussion, and endorsements
- A comprehensive set of impact metrics
- An estimate of Q in percentile terms, e.g., a 90% confidence interval [83, 90]
- A designation of "refereed" that allows the work to "count"

The Cold-Start Problem

- Posting on the Ledger is **no-regrets** for all authors
- A founding board ($N \approx 100$) commits to (a) posting their work and (b) providing N endorsements
- Leading journals add language to their instructions to authors
- Institutional authority lends credibility (Wharton, INFORMS, and so on)
- To start, all existing work could be scraped, indexed, analyzed, and reformatted