The Prestige-Testability Tradeoff in Science

Kurtis A. Hingl*
George Mason University

Job Market Paper

October 27, 2025

[latest version]

Abstract

Where ideas are difficult to test directly, does the scientific community rely more on prestige markers to evaluate them? In this paper, I adopt the cultural evolutionary concept of "prestige," translate it into economics through a simple reputation model, and propose this hypothesis of a prestige-testability tradeoff: scientific fields that are less testable rely more on prestige markers, manifesting a higher concentration. I present empirical evidence of this prestige-testability tradeoff in two ways. Firstly, in bibliographic data of the corpus of scientific research from 1900 to 2015, I find that the concentration of author prestige markers—citations and h-indexes—is consistently negatively associated with a straightforward measure of testability—the incidence of the word "test" in the titles—across nineteen fields and across subfields within each field. Secondly, I use the occurrence of a paradigm shift toward more testability in the mid-1990s as an event study: the "credibility revolution" in microeconomics. Though not truly exogenous, this paradigm shift reflects a testability shock that is suitably uncovered by a staggered event-study design. I find that the credibility revolution administers a leveling effect on its adopters, based on various citation metrics and share of papers in top-five journals: authors below-median pre-adoption on these prestige markers see clear and persistent increases in their prestige markers, while their above-median peers do not, which I interpret as evidence for the prestige-testability tradeoff. I argue that this prestige-testability tradeoff framework is an important lens for viewing the organization of science, an important factor in a number of science policy decisions, and likely a feature of other social learning environments.

JEL Codes: O31, L14, I23
Keywords: economics of science, scientific fields, prestige, testability, reputation, paradigm shift, credibility revolution.

*Ph.D. Candidate, Department of Economics, George Mason University. Email: khingl@gmu.edu. This project was made possible by the Mercatus Center and was supported by resources provided by the Office of Research Computing at George Mason University (https://orc.gmu.edu), funded in part by grants from the National Science Foundation (Award Number 2018631). I thank Yahya Alshamy, Ben Bauer, Peter Boettke, Tyler Cowen, Matthew Kelly, Janna Lu, Caleb Petitt, Jonathan Schulz, Marcus Shera, Alex Tabarrok, and Tegan Truitt for feedback and support. All errors are my own.

> *The beauty of poetry is a matter of such nicety, that a young beginner can scarce ever be certain that he has attained it. Nothing delights him so much, therefore, as the favourable judgments of his friends and of the public; and nothing mortifies him so severely as the contrary... Mathematicians, on the contrary, who may have the most perfect assurance, both of the truth and of the importance of their discoveries, are frequently very indifferent about the reception which they may meet with from the public.*
> Adam Smith (1759, 123-124)

1 Introduction

Economists have long had an interest in the industry of science as a source of economic growth, an existence proof of public goods provision, and an interesting case of labor markets, reward structures, and human capital development (Stephan 1996; Mokyr 2016). But the primary currency in the market place of ideas is prestige, not dollars, euros, or yen.

As defined in cultural evolutionary theory, *prestige* is voluntarily awarded status, followers freely choosing a leader (Henrich and Gil-White 2001). By deferring to and learning from a prestigious role model, a group can transmit information with high-fidelity; this has been shown to lead to group-level adaptive advantages in a variety of settings (Henrich et al. 2015; Henrich 2016). I argue that the industry of science serves as a prototypical example of a domain of social learning organized by prestige and ripe for study. Firstly, the industry is built around the growth of knowledge: researchers present ideas, hypotheses, methods, and tools, and other researchers freely choose to award them status. Secondly, the output of the industry is well recorded in journals and books. Thirdly, the custom of referencing other research via formal and voluntary citation provides a transparent measure of prestige and a natural entry point for observational study. Prestige, as the currency of science, buys a researcher the right to spread their ideas.

Using prestige as an organizing principle in science comes with a number of benefits and costs. On the one hand, science is built by standing on the shoulders of giants: by copying the path of the previously successful, a new researcher advances faster to the knowledge frontier where she can spend her resources building the cumulative stock of knowledge. On the other hand, relying too much on prestige markers can lead to socially wasteful status games, slowing the expansion of the knowledge frontier.[1] With a premium on novelty and priority, researchers compete for fixed prestige rents and

[1] Some of these costs are the result of "prestige bias" as studied in the literature on psychology and cultural evolution (Jiménez and Mesoudi 2019; Egozi and Ram 2024).

race for discovery (Merton 1957; 1961; Hill and Stein 2025a; 2025b).

But the level of prestige deference in science does not exist in a vacuum. My central argument is that the *testability* of the subject of study affects how much the field comes to rely on prestige, and in turn the organizational structure of the field. I propose the following simple hypothesis: this relationship is negative. Scientific fields that are less testable rely more on prestige markers and thus exhibit a higher concentration of prestige markers.

Note that I am making no claims about the desirability of prestige deference in any given case; I claim it is a function of the kind of knowledge being produced in a given field. Again, prestige deference comes with both benefits and costs, as Adam Smith (1759, 123) recognizes, presented in the epigraph above. So to use his example: poetry and mathematics are both valuable fields of knowledge, but the *kinds* of knowledge are different and thus the fields differ in their organization and how much they depend on the opinion of their peers.

Prestige is closely related to the economic concept of *reputation*, used to describe a consumer's assessment of the quality of a seller (Kreps and Wilson 1982; Shapiro 1983; Klein 1997). These terms differ in that prestige is awarded by one's peers, while reputation is awarded by the other side of the market. However, in the case where sellers are also buyers (like researchers in the industry of science) these two concepts overlap. With this motivation, in Section 2 I take the concept of prestige and translate it into economics in a simple reputation model, and I also introduce the concept of testability as a variation in the delay until the quality of a good is revealed. A perfectly testable good is akin to a search good, while a less testable good is akin to an experience or credence good (Nelson 1970; Darby and Karni 1973). The punchline of the model is the following: in order to meet the participation constraint, buyers require less certainty about seller reputation (or prestige markers) if the quality of the good is revealed earlier in time. Consequently, in markets with longer-lived quality uncertainty, the threshold level of seller reputation is higher, leaving a higher concentration of reputation (or prestige markers). By interpreting the model's parameters in the context of science, this serves as a grounding for my hypothesis of the prestige-testability tradeoff.

I test my hypothesis with two methods. Firstly, to get a sense of the universe of the industry of science for the past century, I look at a range of nineteen scientific fields and their subfields (twelve each) from 1900 to 2015, and I use OLS to elicit the association between the concentration of prestige markers and testability. More specifically, for each field and subfield in each year, I calculate the Herfindahl–Hirschman index (HHI) of (a) author citations, (b) paper citations, and (c) author h-index

(each log-ed for interpret-ability). I then regress these on "testability", measured simply as the incidence of the string " test" (space included) in paper names, normalized to a percentile gradient. *I find the concentration of prestige markers is consistently negativity associated with testability, among the nineteen fields and among the subfields in each field.* See Section 3 for details on the regression analysis.

Secondly, and more narrowly, I ask: does a testability shock in a specific field change the dynamics of prestige deference? To answer this question, I study the case of a paradigm shift toward more testability: the "credibility revolution" in microeconomics beginning in the mid-1990s. Specifically, I use a staggered event study design to assess the effects on 3,284 authors who adopt the new paradigm, using not-yet-adopters as the comparison group.

As expected for any successful paradigm shift, authors gain prestige post adoption (Kuhn 1962). But importantly for my question, authors with lower prestige before adoption see bigger effects. That is, *I find that the credibility revolution administers a leveling effect on its adopters, based on five-year citations, ten-year citations, likelihood of scoring a "hit" paper, and share of papers in top-five journals: authors below-median pre-adoption on these prestige markers see clear and persistent increases in their prestige markers, while their above-median peers do not.* While not truly exogenous, the event study design aptly reveals the dynamics of a paradigm shift as a shock to the scientific field.[2] Indeed I hesitate to claim strong causal status of any of my estimates but rather focus on the heterogeneity among the observed effects.[3]

Related Literature. Along with the literature on prestige and reputation, this paper contributes to the growing literature on the economics of science, beginning with Smith (1759), reignited by Nelson (1959) and Arrow (1962), and summarized by Stephan (1996). Polanyi (1962) and Tullock (1966) model the scientific enterprise as a set of rules by which researchers interact, jointly building the broader organization of science—these I take as influential in my model. Regarding prestige in science (more loosely defined), the literature has noticed the inequality in influence since at least Robert Merton (1968), who famously calls this the *Matthew Effect*: "the rich get richer, the poor get poorer." In economics, this inequality is often called a winner-take-all market (when the mechanism is supply-side: Cook and Frank 2010; 2013) or a positional good (when the mechanism is demand-side: Carlsson et al. 2007; Schneider 2007). Regarding my methods, many recent papers tackle economic questions in science through the use of bibliographic data, and from these I draw inspiration (Wu et al. 2019;

[2] See Azoulay et al. (2019) who look at the shock of an unexpected death of a star researcher for a similar difference-in-differences empirical design.

[3] See Appendix C.1 for discussion about appropriate causal interpretations.

Azoulay et al. 2019; Angrist et al. 2020; Liu et al. 2023; Hager et al. 2024; Hill and Stein 2025a; 2025b; Hill et al. 2025; Tripodi et al. 2025). Finally, Huber et al. (2022) ask a question closely related to my hypothesis: does the prestige of the listed author names affect peer-review feedback? The authors ran a preregistered field experiment soliciting 534 peer reviews on a finance paper while only listing either economics Nobel Laureate Vernon Smith or his (relatively unknown) coauthor Sabiou Inoua; they find significantly lower rejection rates and better comments across the board with the name of the Nobel listed instead of the novice.

This paper combines the narrowly economic approach with the approach of the cross-disciplinary "science of science" papers: that is, I propose a model-grounded hypothesis and test it using the micro-econometric toolkit, while the research question explicitly compares characteristics across scientific fields. My unique contribution is twofold: first, it is to translate the concept of prestige into the language of economics through reputation, and second, it is to use this to examine the organization of science based on two substitutable methods of learning. I aim to uncover one aspect of the scientific "rules of the game," and if I am successful, this has the potential to inform downstream questions about incentive structures, evaluation funding, and allocation of researchers.

Data. The data used for all analyses in this paper are from SciSciNet (Lin et al. 2023) a large-scale relational data lake of scientific contributions, authors, outlets, and institutions. SciSciNet builds on Microsoft Academic Graph (Sinha et al. 2015; Wang et al. 2019; Wang et al. 2020) and OpenAlex (Priem et al. 2022) and has become a standard for open and transparent research on the science of science. I limit my analyses to the years 1900 to 2015. See Appendix B for relevant summary statistics.

The paper proceeds as follows: I present the model in Section 2, conduct the wide-scale regression analysis in Section 3, conduct the paradigm shift event study in Section 4 and conclude with a discussion in Section 5.

2 A Model of Prestige-Testability Tradeoff

2.1 General model

Let us begin with a simple reputation model. Suppose there are two markets A and B that are independent and are identical except for one distinguishing feature, a delay in quality revelation. At time $t = 0$, the good is produced and sold in both markets. The quality of good X_A is revealed at time $t = 1$ as either high or low (h or l), but the quality of good X_B is revealed at time $t = 2$.

For simplicity, buyers are only interested in buying the high-quality good X_i^h at price p^h, but there is some chance the seller cheats and that they are sold a low-quality good at the high-quality price. If a seller cheats, this imposes a per-period penalty c on the buyer for every period he holds the good. That is, in market A, when the buyer finds out he bought a low quality good at time $t = 1$, he incurs $c_A = c > 0$. In market B, when the buyer finds out he bought a low quality good at time $t = 2$, he incurs $c_B = c + \beta c = c(1 + \beta)$ where $\beta \in [0, 1]$ is the one-period discount factor.

Finally, let α denote the buyer's belief that a given seller will not cheat on the next sale. This can (but needn't) be imputed from past play: if the buyer observes the fraction of times each firm has cheated in the past $1 - \alpha$, and thus infers that an X_i^h is truly high quality with probability $\alpha \in [0, 1]$.

The expected utility of a buyer in market i is thus

$$U_i(\alpha_i) = \alpha_i(v - p^h) + (1 - \alpha_i)[-c_i - p^h], \tag{1}$$

where v is the value he derives.

The buyer's participation constraint is thus $U_i(\alpha_i) \geq 0$:

$$\alpha_i(v - p^h) + (1 - \alpha_i)[-c_i - p^h] \geq 0. \tag{2}$$

Solve for α_i:

$$\alpha_i \geq \frac{c_i + p^h}{(v - p^h) + (c_i + p^h)} = \frac{c_i + p^h}{v - c_i}. \tag{3}$$

.

Since the cost of being cheated in market B carries over into two periods, $c_A < c_B$, and thus $\alpha_A^* < \alpha_B^*$. Or, more specifically,

$$\alpha_A^* = \frac{c + p^h}{v - c} < \frac{c(1 + \beta) + p^h}{v - c(1 + \beta)} = \alpha_B^* \tag{4}$$

This has the following intuitive interpretation:

Proposition 1. *Consumers require more quality assurance in a market with longer-lived quality uncertainty.*

In other words, a market where a good's quality remains uncertain for longer will rely more on reputation (in this case past performance). We can generalize this based on variations in the discount factor β and delay in quality revelation. For the myopic buyer with $\beta = 0$, $c(1 + \beta) = c$ and thus $\alpha_A^* = \alpha_B^*$; reputation does not matter for him. For a new market j where the quality is revealed after time $t = \tau > 2$, $c_j = c(1 + \beta)^\tau$, making the reliance on reputation stronger.

Where α_i corresponds to a seller's reputation $\hat{\alpha}_i \in [0, 1]$ as the past non-cheat-rate, markets A and B can be compared directly on concentration of $\hat{\alpha}_i$. Assume the baseline CDF for the sellers'

reputations is $G(x)$, and buyers screen using cutoffs α_A^* and α_B^* with $\hat{\alpha}_A^* < \hat{\alpha}_B^*$ by Inequality 4. The distribution of reputations among *active* sellers in market i is the conditional CDF

$$G_i(x) \equiv \Pr(\hat{\alpha} \leq x \mid \hat{\alpha} \geq \alpha_i^*) = \frac{G(x) - G(\alpha_i^*)}{1 - G(\alpha_i^*)} \quad (x \geq \alpha_i^*). \tag{5}$$

Because G is nondecreasing and $\alpha_A^* < \alpha_B^*$, it follows that

$$G_B(x) \leq G_A(x) \tag{6}$$

for all x, and with strict inequality on $[\alpha_B^*, 1)$ when $G(x) < 1$ (proof in Appendix A). Thus the active reputation distribution in B first-order stochastically dominates that in A: slower revelation raises the participation cutoff and truncates the active pool to the right, leaving only higher-reputation sellers. This leads to the straightforward relationship between length of uncertainty, reliance on reputation, and equilibrium of reputation-level concentration based on the participation constraint:

Proposition 2. *When consumers' quality assurance beliefs correspond with past seller performance through a reputation marker, markets with longer-lived quality uncertainty show greater concentration of reputation markers.*

2.2 The model in research production

In the context of scientific research production, Proposition 1 can be restated as follows:

Proposition 3. *Science evaluators require more quality assurance in a scientific field with longer-lived quality uncertainty.*

Fields that produce ideas that are more testable reduce the quality revelation time, and thus we should expect these fields to rely less on the the reputation of the researcher in the evaluation of the ideas.

Likewise, Proposition 2 can be restated as follows:

Proposition 4. *When science evaluators' quality assurance beliefs correspond with past researcher performance through a reputation marker, fields with longer-lived quality assurance show greater concentration of reputation markers.*

For completeness, let us interpret each parameter in the domain of scientific production. The good is an article, book, or otherwise one unit of scientific output; the producing firm is the author(s) of the work; the buyer is the scientific community. Reputation based on success in the past can be directly interpreted as prestige.[4] Past citations accrued to an author is thus a natural, albeit imperfect,

[4]Reputation based on an extra-market signal would not be a measure of prestige, conventionally defined (Henrich and Gil-White 2001).

measure of prestige.[5] Other measures such as productivity, h-index, or institution ranking are also natural proxies, but note that while prestige is a powerful mechanism, it is a nebulous term, and is best defined in context.

In practice, the value v placed on a unit of output, the price p^h, and even the cost c of holding a low quality good may all vary across scientific fields, corroborating our ability to directly compare prestige-reliance and concentrations across fields. For example, a new output in chemistry (say a drug) may be more valuable than a new output in history ($v_A > v_B$), the price the community is willing to pay may reflect this ($p^h_A > p^h_B$), and we may suffer less from a mistaken historical account than from an inappropriately prescribed drug ($c_B < c_A$).

Notice in Inequality 3, however, that as the limit of $(v - p^h) \to 0$, $\alpha \to 1$. This means that with no consumer surplus, the buyer must be certain he is receiving a high quality good and will only buy from a firm with no past low quality output. While it is logically possible that there is more consumer surplus per unit in certain fields of science, there is no a apriori reason to expect this to be connected with testability nor consistent over time. To make the general model more tractable for science, we can assume that the price paid and value received are non-monetary for the buyer—whether or not the community approves of a new idea is constant in shadow price across disciplines. This matches the common sentiment and empirical findings that the production of specific scientific inquiry is largely independent of financial incentives (Stern 2004; Myers 2020).

To summarize the model, the scientific community accepts or rejects ideas from researchers based on expected utility, which is a function of a researcher reputation belief and the costs of adopting a low-quality idea. In other words, there is some combination of researcher prestige and idea testability that together allow the community to accept the idea; those attributes are substitutes in consumption. In fields that are less testable—fields where ideas have longer-lived quality uncertainty—a higher reputation threshold is necessary to satisfy the participation constraint, and this mechanically leads to greater concentration of prestige markers. With the above qualifications, this grounds my simple hypothesis of a prestige-testability tradeoff in science.

[5] Unlike in Shapiro (1983), researcher reputation cannot be transferred—it expires with death. This gives the straightforward prediction that researchers will "cash in" on their reputation and produce less testable research as they reach some age threshold. I leave this for further research.

3 Regression analysis

My first empirical test takes the widest possible lens. Namely, I look for an association between the testability of a field of study and the concentration of its prestige markers. To this end, I adopt the universe of bibliographic data from nineteen scientific fields from 1900 to 2015 from SciSciNet (Lin et al. 2023).[6] This gives me 149,967,677 distinct authors and 74,013,927 contributions published as journal articles (for full summary statistics broken down by field, see Appendix B). For these associative estimates, I use the following three measures as prestige markers: author citations, author h-index, and paper citations. Because these are voluntarily awarded by the scientific community, I expect this to capture a realistic portrait of the prestige hierarchies, even if imperfect.

In the first series of OLS regressions I take the scientific field in each year as the unit of analysis, testability as the independent variable, with the concentration (HHI) of the three prestige measures as the dependent variables.

$$log\ \text{HHI}_{f,t} = \alpha + \beta\, \text{Testability}_{f,t} + \mathbf{X}_{f,t}\gamma + \lambda_t + \varepsilon_{f,t} \tag{7}$$

where f is field, t is year, $\mathbf{X}$ is controls, and λ is year fixed effects. In the second series of OLS regressions, I repeat the procedure, but take the scientific subfield in each year as the unit of analysis and examine variation within the fields by adding μ as a field fixed effect.

$$log\ \text{HHI}_{s,t} = \alpha + \beta\, \text{Testability}_{s,t} + \mathbf{X}_{s,t}\gamma + \mu_f + \lambda_t + \varepsilon_{s,t} \tag{8}$$

The coefficient of interest is β, which I hypothesize to be negative. Table 1 presents a description of each of unit and variable including controls, and Table B1 presents summary statistics, and Table 2 presents results. Across all regressions for fields and subfields, the coefficient of interest is indeed negative. As an interpretation of a typical coefficient: a 10 percentile increase in testability across fields is associated with a 6.85% decrease in the concentration of Author Citations, holding year fixed (Table 2, Panel A, column 3), or a 10 percentile increase in testability across subfields is associated with an 10.15% decrease in the concentration of Author H-index, holding year and field fixed (Table 2, Panel B, column 6). Figure 1 shows the testability gradient mapped on to Author Citation HHI; this corresponds to the first column in Table 2, Panel A, though it is not the regression proper.

[6] Each paper is assigned one of these nineteen "top" fields in SciSciNet: Art, Biology, Business, Chemistry, Computer science, Economics, Engineering, Environmental science, Geography, Geology, History, Materials science, Mathematics, Medicine, Philosophy, Physics, Political science, Psychology, Sociology.

Table 1: Variables and Definitions for OLS Analysis

Item	**Type**	**Definition**
Field	Unit for 1.1-1.9	Unique top `fieldid` (19 total), assigned by `SciSciNet`.
Subfield	Unit for 2.1-2.9	Unique sub `fieldid` mapped to a top field using the 12 most common per year.
Testability	Independent	Percent incidence of the string " test" (with leading space) in titles in each unit–year (random 10% sample for fields and subfields), normalized to a percentile.
Paper-citation HHI	Dependent	Herfindahl–Hirschman Index over all paper citations in the unit–year. Logged for interpretation.
Author citation HHI	Dependent	Herfindahl–Hirschman Index over all author citations in the unit–year. Logged for interpretation.
Author h-index HHI	Dependent	Herfindahl–Hirschman Index over all lifetime h-indexes for authors active in the unit–year. Logged for interpretation.
Team size	Control	Median team size calculated over unit-year.
# Active authors	Control	Number of active authors calculated over unit-year.

Note. See Table B1 in Appendix B for summary statistics.

Table 2: OLS Results: Testability on Concentration of Prestige Markers

Panel A: Fields

	Dependent variable:								
	log Author-Cite HHI			*log* Author H-index HHI			*log* Paper Cite HHI		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Testability (percentile)	−1.003***	−0.691***	−0.685***	−0.941***	−0.658***	−0.646***	−0.811***	−0.548***	−0.540***
	(0.111)	(0.091)	(0.091)	(0.086)	(0.064)	(0.064)	(0.104)	(0.089)	(0.089)
Team Size		−1.520***	−1.427***		−1.377***	−1.177***		−1.278***	−1.143***
		(0.047)	(0.064)		(0.033)	(0.044)		(0.046)	(0.062)
# Active Authors			−0.000**			−0.000***			−0.000***
			(0.000)			(0.000)			(0.000)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,204	2,204	2,204	2,204	2,204	2,204	2,204	2,204	2,204
R^2	0.643	0.762	0.762	0.784	0.882	0.884	0.628	0.729	0.730
Adjusted R^2	0.623	0.748	0.749	0.772	0.875	0.878	0.608	0.713	0.715

Panel B: Subfields

	Dependent variable:								
	log Author-Cite HHI			*log* Author Hindex HHI			*log* Paper Cite HHI		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Testability (percentile)	−0.984***	−0.848***	−0.852***	−1.145***	−1.003***	−1.015***	−0.832***	−0.740***	−0.747***
	(0.031)	(0.029)	(0.029)	(0.027)	(0.025)	(0.025)	(0.029)	(0.028)	(0.028)
Team Size		−0.612***	−0.595***		−0.635***	−0.574***		−0.413***	−0.379***
		(0.011)	(0.012)		(0.009)	(0.011)		(0.011)	(0.012)
# Active Authors			−0.000***			−0.000***			−0.000***
			(0.000)			(0.000)			(0.000)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	24,899	24,899	24,899	25,054	25,054	25,054	24,899	24,899	24,899
R^2	0.833	0.852	0.853	0.887	0.905	0.906	0.807	0.818	0.818
Adjusted R^2	0.832	0.852	0.852	0.886	0.904	0.905	0.806	0.817	0.817

Note. There is a negative association between testability and these three prestige markers, across the nineteen fields, and across subfields, holding fields fixed. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

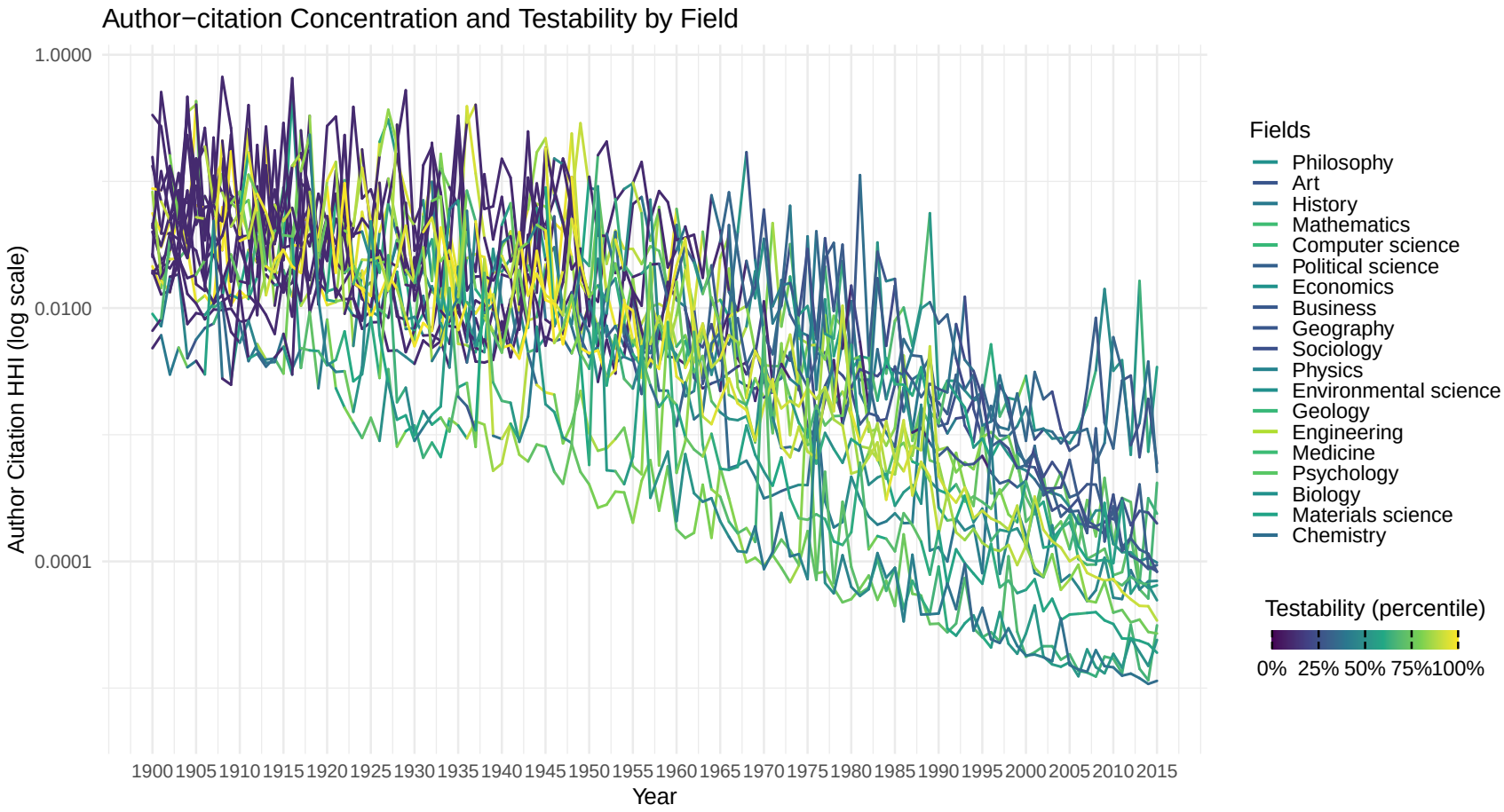


Figure 1: Author-citation Concentration and Testability. *Note:* The prestige-testability tradeoff hypothesis says that there should be lower testability (dark) where there is higher concentration of prestige markers (top), but note that all regressions reported in Table 2 include year fixed effects. Testability is measured as the incidence of the string " test" (space included) in titles among the fields, and factored as percentile. HHI is calculated as $hhi_{f,t} = \sum_{i=1}^{n} \left(\frac{cites_{i,t}}{totalcites_{f,t}} \right)^2$, where f is a field, t is a year, i is an individual author who publishes a contribution in that field-year, n is total active authors in that field-year. Number of field-years = 2204. The order of fields in the key is ranking in last observed year (2015).

4 Paradigm Shift Event Study

Edward Leamer (1983) famously wrote the following in a paper entitled "Let's Take the Con out of Econometrics": "Hardly anyone takes data analyses seriously. Or perhaps more accurately, hardly anyone takes anyone else's data analyses seriously. Like elaborately plumed birds who have long since lost the ability to procreate but not the desire, we preen and strut and display our t-values." In 2010, Joshua Angrist and Jörn-Steffen Pishke (2010, 3-4) reflect: "[Leamer's] critique had a refreshing emperor's-new-clothes earthiness that we savored on first reading and still enjoy today. But we're happy to report that Leamer's complaint that 'hardly anyone takes anyone else's data analysis seriously' no longer seems justified. Empirical microeconomics has experienced a credibility revolution, with a consequent increase in policy relevance and scientific impact." In 2021, Angrist was awarded the Nobel Prize in Economics with David Card and Guido Imbens for their contributions that sparked this "credibility revolution" in empirical economics; their main contributions were developed in the early- and mid-1990s.

The credibility revolution is a clear example of a scientific paradigm that aims to shift a field toward "testability." For the sake of this paper, I am agnostic to the metaphysical truth claims of any given study, within the credibility revolution or otherwise. But it is not my interpretation that matters: as long as the new scientific practices revolutionized the ability for practitioners to "take anyone else's data analyses seriously," we have a genuine shift in the testability parameter.[7]

Here I take on the task of parsing out the effect that this paradigm shift had on prestige of its participants. That is, I am interested on the effects of researchers who select into the credibility revolution. With any successful paradigm shift, testability-related or otherwise, we would expect the participants to gain prestige. But if my testability-prestige tradeoff hypothesis holds, we should expect the strongest effects for those who did not have prestige prior to their participation in a "more" testable subject matter. Intuitively a shift toward testability can lift up the previously unrecognized—the quality of the output is revealed much sooner.

To approach my measurement, I adopt a staggered event study design. The pool of researchers is 3,284 authors who publish a paper that cites a set of five "credibility revolution" seed papers: Angrist and Imbens (1994), Card and Krueger (1994), Bound et al. (1995), Angrist et al. (1996), and Staiger and Stock (1997) (see Appendix C.2 for summary statistics). I take an author's first publication

[7] In terms of my model in Section 2 this is a shift from a later quality revelation time to an earlier quality revelation time.

that cites a seed paper as the method of treatment, and once a researcher is treated they remain forever treated. While this is perhaps not the tightest possible definition of joining the credibility revolution—an author could critically cite a seed paper—I use this method for transparency and to be consistent in using citations as a measure of prestige-deference. This setup fits nicely into the event study method proposed by Sun and Abraham (2021): it reflects staggered treatment and we certainly expect heterogeneous treatment effects. The main specification is as follows:

$$y_{it} = \sum_{k \in \mathbb{Z} \setminus \{-1\}} \beta_k \underbrace{\left(\sum_{g \in \mathcal{G}} \mathbf{1}\{C_i = g\} \, \mathbf{1}\{t - g = k\} \right)}_{\text{Sun–Abraham cohort} \times \text{event-time dummies}} + \mu_t + \varepsilon_{it} \, ,$$

where outcomes and variables are defined in Table 3. Note that this is built from a simple 2×2 difference-in-differences design. At its core, the goal is to compare authors who adopt the new "testability" regime to their potential outcome had they not adopted the regime, and then ultimately how this "treatment effect" varies among the kinds of authors who participate (see Appendix C.1 for a full interpretation based on potential outcome notation). Because the sample is restricted to only authors who eventually adopt the new paradigm, the source of variation is the timing of adoption.

Along with this main specification I also estimate a "strict" version where I introduce subfield level fixed effects (along with year) $\mu_{f(i),t}$. Finally, I investigate the heterogeneity by below and above median pre-adoption prestige.[8] While the actual outcomes of adopting authors may be interesting in its own right, this heterogeneity is the relevant test of my hypothesis.

Results. The main event study paths are presented in Figure 2 and the event study paths broken down by pre-treatment heterogeneity are presented in Figure 3. In the aggregate, there is a modest increase in all outcomes post adoption. Figure 2, A shows that post-treatment, cohorts receive between about 0.1 to 0.2 standard deviations more citations within five years of publishing their papers, compared to the estimated counterfactual of the not-yet-treated; this effect persists over the decade post treatment. Similarly, Figure 2, B shows an increase in about 0.1 to 0.3 standard deviations more citations within 10 years of publication, compared to the counterfactual. Figure 2, C shows that treated cohorts are about 3% to 8% more likely to have a "hit" paper in their field, defined as being in the top 10 % of papers in lifetime citations. Figure 2, D shows that treated cohorts publish Top-5 papers 1% to 3% more than the control group, though this effect vanishes with subfield controls. For

[8] Or: $y_{it} = \sum_{k \in \mathbb{Z} \setminus \{-1\}} \beta_k^{(q)} \left(\sum_{g \in \mathcal{G}} \mathbf{1}\{C_i = g\} \, \mathbf{1}\{t - g = k\} \right) + \mu_t^{(q)} + \varepsilon_{it}, \qquad i \in Q_q.$

Table 3: Variables and Definitions for Event Study: Main: Strict, and Heterogeneity

Item	Symbol	Definition
Unit of observation	i, t	Author i in calendar year t.
Treatment cohort (adoption year)	C_i	First year i publishes a paper that cites any of the five seed papers (once treated, always treated).
Event time (reference $k = -1$)	$k \equiv t - C_i$	Relative year to adoption; $k = -1$ is the omitted (reference) period.
Outcomes	y_{it}	Field-normalized 5 year citations (c5_z), field-normalized 10 year citations (c10_z), top 10 % paper in the field by lifetime citations (hit10), top-5% share (top5).
SA cohort × event-time dummies	SA_{itk}	Sun–Abraham (2021) basis: $SA_{itk} \equiv \sum_{g \in \mathcal{G}} \mathbf{1}\{C_i = g\}\,\mathbf{1}\{t - g = k\}$.
Interaction weight	β_{gk}	Sun–Abraham interaction weight for cohort g at event time k; event-time coefficients are aggregated as $\hat{\beta}_k = \sum_{g \in \mathcal{G}_k} \beta_{gk}, \tau_g(k)$ with $\sum_{g \in \mathcal{G}_k} \beta_{gk} = 1$ for each k. This allows cohorts to contribute based on size without negative weighting.
Main fixed effects	μ_t	Calendar-year FE (absorbed).
Strict fixed effects	$\mu_{s(i),t}$	Subfield × year FE (absorbed), where $s(i)$ maps author i to subfield-year handles.
Heterogeneity, pre-period set	$\mathcal{T}_i^{\text{pre}}$	Author-specific pre-treatment years observed: $\mathcal{T}_i^{\text{pre}} = \{t < C_i : \ y_{it} \text{ observed}\}$.
Pre-treatment summary for ranking	$\bar{y}_i^{\text{pre}}$	Mean (or median) outcome over $\mathcal{T}_i^{\text{pre}}$ used to rank authors prior to treatment for that outcome.
Heterogeneity group	$Q_i \in \{1, 2\}$	Two heterogeneity bins for median split of $\bar{y}_i^{\text{pre}}$ among not-yet-treated authors in the risk set.
Group indicator(s)	$D_{iq} = \mathbf{1}\{Q_i = q\}$	Time-invariant dummies; interactions $SA_{itk} \times D_{iq}$ allow the path to differ by pre-period heterogeneity.
Event study estimating equations	—	*Main/Strict*: $y_{it} = \sum_{k \neq -1} \beta_k \, SA_{itk} + \text{FE} + \varepsilon_{it}$. *Heterogeneity*: $y_{it} = \sum_{k \neq -1} \beta_k \, SA_{itk} + \sum_{k \neq -1} \delta_k \, SA_{itk} D_{i2} + \text{FE} + \varepsilon_{it}$.
Inference (clusters)	—	Two-way clustering by author and FE dimension (year for Main; Subfield ×year for Strict).
Figure trimming	$k \in [-15, 15]$	For plots, event-time paths are trimmed to a symmetric window where support is adequate.

each outcome, pre-trends are flat, except for a jump within 2-5 years of treatment, arguably due to anticipatory behavior. Anticipation may be likely in this case, as papers take time to achieve final publication, but here I present an unadjusted event-study for transparency.[9] These results, though

[9] See Appendix C for a full discussion of assumptions including around no-anticipation.

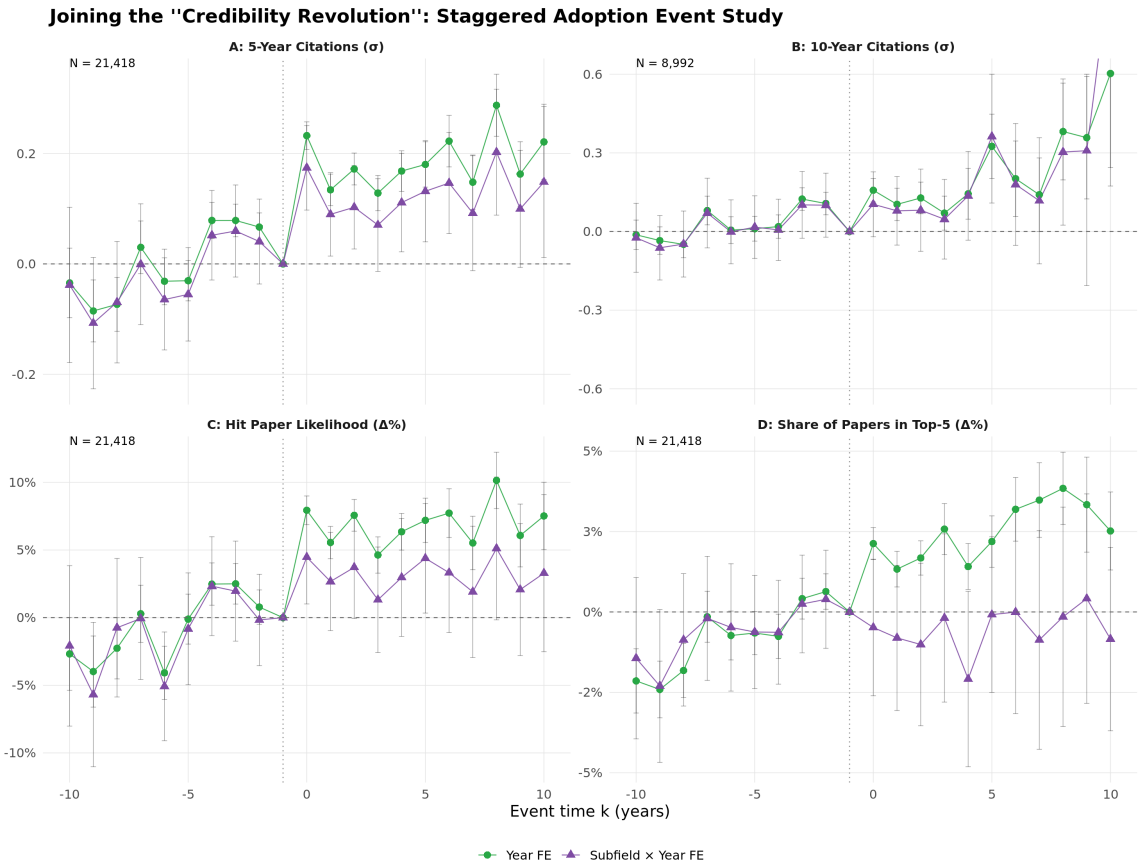


Figure 2: Authors who cite credibility revolution seed papers see the following outcomes: **Panel A** 0.1 to 0.2 standard deviations increase in their 5-year citations, sustained for 10 years post adoption; **Panel B** 0.1 to 0.3 standard deviations increase in their 10-year citations, sustained for 10 years post adoption; **Panel C** 3% to 8% increase in their likelihood of achieving a "hit paper" (defined as a paper achieving lifetime citations in the top decile of the field); **Panel D** a 2% to 3% increase in the share of their papers that make it into a Top-5 journal. *Controlling for subfields, these effects are not all statistically significant.* Staggered event study paths are calculated using Sun and Abraham (2021) interaction weighted estimates; units are adoption-year cohorts, reported N is author-years. Reference year is $k = -1$. See Appendix C.2 for summary statistics.

perhaps real, say little with respect to my prestige-testability hypothesis, however.

To shed light on the hypothesis of interest, I break the event study estimates down by the heterogeneity of the outcomes of interest in the *pre*-treatment period. For this exercise, I calculate each author's average pre-treatment outcome and slit the sample on a simple binary below and above the median. My hypothesis would imply that lower-prestige individuals benefit more from the testability shock. Figure 3 give us visual credence.

Figure 3 shows that the low-baseline group consistently sees greater post adoption effect compared to their high-baseline peers, and in many cases the the high-baseline group sees null or negative effects. In particular, low-prestige authors see a 0.1 to 0.3 standard deviations increase in their 5-year citations,

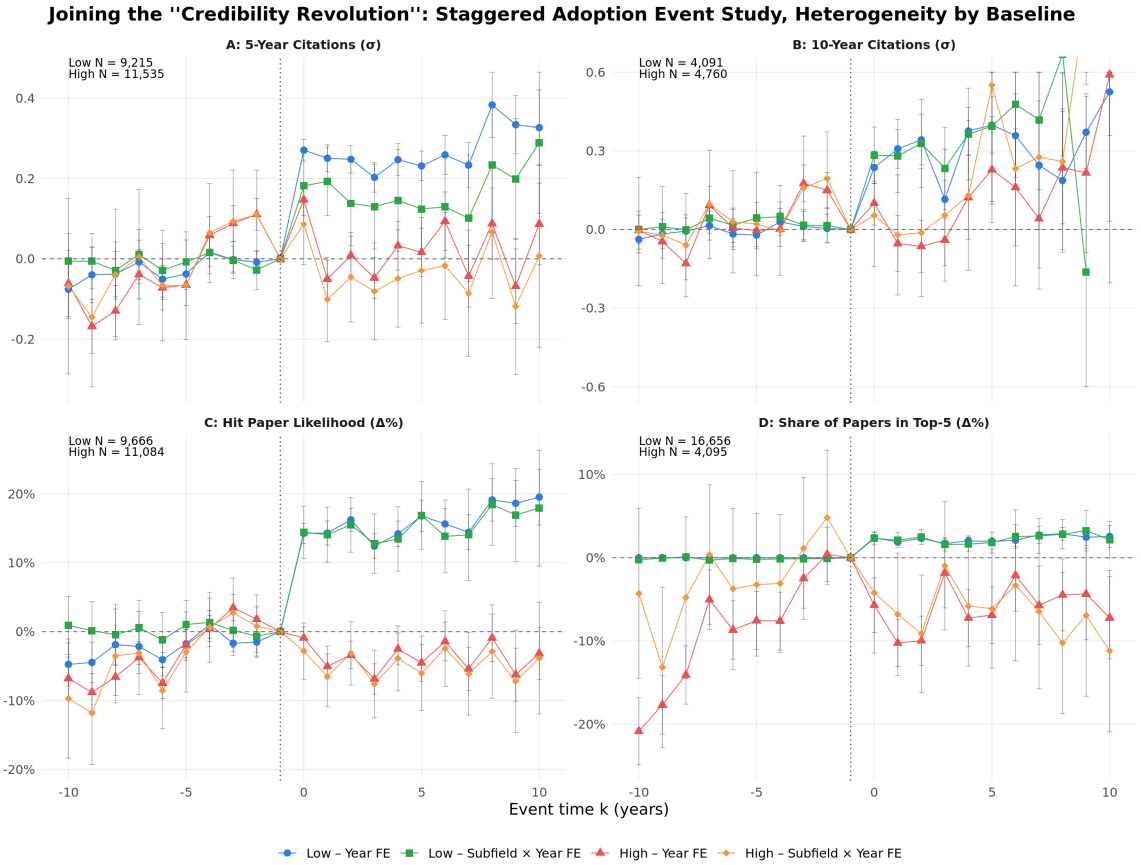


Figure 3: Unpacking by pre-adoption heterogeneity, the gains by joining the credibility revolution are notably concentrated in the low-prestige authors. Specifically, authors who cite credibility revolution seed papers see the following outcomes: **Panel A** low-prestige authors see a 0.1 to 0.3 standard deviations increase in their 5-year citations, while their high-prestige peers see a change of −0.1 to 0.1 standard deviations, and this difference is sustained for 10 years post adoption; **Panel B** low-prestige authors see a 0.1 to 0.4 standard deviations increase in their 10-year citations, while their high-prestige peers see a change of −0.1 to 0.2 standard deviations, and this difference is sustained for 5 years post adoption; **Panel C** low-prestige authors see a 12% to 19% increase in their likelihood of achieving a "hit paper" (defined as a paper achieving lifetime citations in the top decile of the field), while their high-prestige peers see a 1% to 7% decrease in their likelihood, and this difference is sustained for 10 years post adoption; **Panel D** low-prestige authors see a 2% to 3% increase in the share of their papers that make it into a Top-5 journal, while their high-prestige peers see a 1% to 10% decrease (note that negative pre-trends for the high-prestige group indicates that their Top-5 share *increased* each year until peaking at $k = -1$). Staggered event study paths are calculated using Sun and Abraham (2021) interaction weighted estimates; units are adoption-year cohorts, reported N is author-years. Reference year is $k = -1$. See Appendix C.2 for summary statistics.

while their high-prestige peers see a change of −0.1 to 0.1 standard deviations, and this difference is sustained for 10 years post adoption; low-prestige authors see a 0.1 to 0.4 standard deviations increase in their 10-year citations, while their high-prestige peers see a change of −0.1 to 0.2 standard deviations, and this difference is sustained for 5 years post adoption; low-prestige authors see a 12% to 19% increase in their likelihood of achieving a "hit paper," while their high-prestige peers see a

1% to 7% decrease in their likelihood, and this difference is sustained for 10 years post adoption; and low-prestige authors see a 2% to 3% increase in the share of their papers that make it into a Top-5 journal, while their high-prestige peers see a 1% to 10% decrease (note that negative pre-trends for the high-prestige group indicates that their Top-5 share *increased* each year until peaking at $k = -1$). I interpret this leveling effect as support for the prestige-testability tradeoff.

5 Conclusion

"Standing on the shoulders of giants" is a necessary step for scientific progress. But, to extend the metaphor, the need for giants is endogenous to the height of the tree of knowledge where the researcher is looking for fruit. This has been my hypothesis. Or: the kind of knowledge in some fields lends itself to be more testable than in others, and thus a variation of the need for prestige deference across fields. This materializes into a higher concentration of prestige markers, which is consistent with the wide-scale evidence I present in Section 3.

Furthermore, changes in the testability of a field can in turn affect the degree to which its prestige hierarchy is maintained. I explore this using an event study design over the credibility revelation, a paradigm shift in economics. I find that pre-adoption low-prestige researchers who opt in to the new paradigm see consistently larger gains than their high-prestige peers. I take this leveling effect as evidence of the prestige-testability tradeoff.

Note that my primary goal has been to elucidate an aspect of the structure of scientific inquiry, and not to promote policy prescriptions. That said, if the hypothesis holds water, it may have implications for other important questions in the organization of science. Here I discuss a few.

1. Reliance on prestige markers to evaluate researchers is not inherently good or bad. Indeed, when the kind of knowledge being produced does not allow for quick and easy test, verification, or falsification we still must find some way to evaluate ideas. In these cases prestige markers may be as good as any.

2. It may be beneficial for society to promote more diversity of scientific ideas than ones that appear through "normal" channels. After all, breakthrough scientific ideas can exhibit huge positive spillover effects, and it has been shown that science, like other ventures, follows a positive risk-reward path (Azoulay et al. 2011; Azoulay and Greenblatt 2025). On this front, the framework from the prestige testability trade-off hypothesis implies different things for different

fields. In the more testable fields and subfields, research evaluators should allow many more shots on goal, reduce their reliance on prestige markers, and allow the scientific community to assess the testable claims quickly. In less testable fields and subfields, research evaluators might have no other choice than to rely on prestige markers for evaluation—that is okay, and it may be futile or destructive to introduce testability criterion on a subject matter that does not warrant it. But it does not follow that evaluators should be limited to judgment based only on publication numbers, citations, or institutional rankings. Prizes and tournaments could reward prestige while reducing the "time to build" reputation, increasing diversity of ideas and turnover of dominant influence.

3. While a testability shock can lead to less reliance on prestige markers and thus constitute a move to a more meritocratic system, testability shocks are not a choice variable of policy makers in government, academia, or scientific publishing; rather, they are endogenous to the research production itself. A paradigm shift that leads to more testability (as with the credibility revolution) may be attractive to the less prestigious, but it must first play by the rules of the incumbent paradigm to be successful. Research evaluators should be mindful of the margins they can control.

4. Testability is only one dimension of the differences in subject matter across scientific fields, and I have avoided discussions of researcher-side field self-selection based on testability or prestige hierarchies. I would argue that any policy that changes testability criteria or prestige marker deference must account for the framework I present, but it is certainly not the only spring from which streams of incentives flow.

5. The prestige-testability tradeoff warrants further consideration as a factor in other aspects of the structure of science. For example, it has been noted that the age of researchers at which major scientific discovers occur varies considerably over time, and varies somewhat across fields (Jones 2009; 2010; Jones and Weinberg 2011). Age of discovery may be a function of how various fields (and time periods) test output which in turn shapes the underlying prestige hierarchy.[10]

As I present here, the prestige-testability tradeoff may prove most fruitful in uncovering the on-the-ground dynamics of paradigm shifts across fields, space, and time. Indeed, given the simplicity of the hypothesis, it likely a feature of a wide variety of social learning environments besides science: will a

[10] In these discussions, however, little attention is paid to the sheer increase in number of researchers since 1950 (see Figure B6).

flood of artificially generated content suppress the direct "testability" in media and spur consumers to require signals of prestige from the presenter? Or conversely, if prestigious individuals abuse their platforms, can this induce a shift toward testability? If the prestige-testability tradeoff is a real structural phenomenon, it may have implications across education, politics, and culture.

References

Angrist, Joshua D. and Guido W. Imbens (1994). "Identification and Estimation of Local Average Treatment Effects". *Econometrica* 62.2, pp. 467–475. DOI: 10.2307/2951620.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables". *Journal of the American Statistical Association* 91.434, pp. 444–455. DOI: 10.1080/01621459.1996.10476902.

Angrist, Joshua D. and Jörn-Steffen Pischke (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics". *Journal of Economic Perspectives* 24.2, pp. 3–30.

Angrist, Joshua D. et al. (2020). "Inside Job or Deep Impact? Extramural Citations and the Influence of Economic Scholarship". *Journal of Economic Literature* 58.1, pp. 3–52. DOI: 10.1257/jel.20181508.

Arrow, Kenneth J. (1962). "Economic Welfare and the Allocation of Resources for Invention". In: *The Rate and Direction of Inventive Activity: Economic and Social Factors*. Ed. by Richard R. Nelson. Princeton, NJ: Princeton University Press. pp. 609–626. DOI: 10.1515/9781400879762-024.

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin (2019). "Does Science Advance One Funeral at a Time?" *American Economic Review* 109.8, pp. 2889–2920. DOI: 10.1257/aer.20161574.

Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso (2011). "Incentives and Creativity: Evidence from the Academic Life Sciences". *RAND Journal of Economics* 42.3, pp. 527–554.

Azoulay, Pierre and Wesley H. Greenblatt (2025). *Does Peer Review Penalize Scientific Risk Taking? Evidence from NIH Grant Renewals*. Tech. rep. NBER Working Paper. Cambridge, MA: National Bureau of Economic Research.

Bound, John, David A. Jaeger, and Regina M. Baker (1995). "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory

Variable Is Weak". *Journal of the American Statistical Association* 90.430, pp. 443–450. DOI: 10.1080/01621459.1995.10476536.

Card, David and Alan B. Krueger (1994). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania". *American Economic Review* 84.4, pp. 772–793.

Carlsson, Fredrik, Olof Johansson-Stenman, and Peter Martinsson (2007). "Do You Enjoy Having More than Others? Survey Evidence of Positional Goods". *Economica* 74.296, pp. 586–598.

Cook, Philip J. and Robert H. Frank (2010). *The Winner-Take-All Society: Why the Few at the Top Get So Much More than the Rest of Us*. New York: Random House.

Darby, Michael R. and Edi Karni (1973). "Free Competition and the Optimal Amount of Fraud". *Journal of Law and Economics* 16.1, pp. 67–88.

Egozi, Saar and Yoav Ram (2024). "Prestige Bias in Cultural Evolutionary Dynamics". *Royal Society Open Science* 11.7, p. 230650.

Frank, Robert H. and Philip J. Cook (2013). "Winner-Take-All Markets". *Studies in Microeconomics* 1.2, pp. 131–154.

Hager, Sebastian, Carlo Schwarz, and Fabian Waldinger (2024). "Measuring Science: Performance Metrics and the Allocation of Talent". *American Economic Review* 114.12, pp. 4052–4090. DOI: 10.1257/aer.20230515.

Henrich, Joseph (2016). *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Henrich, Joseph, Maciej Chudek, and Robert Boyd (2015). "The Big Man Mechanism: How Prestige Fosters Cooperation and Creates Prosocial Leaders". *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1683, p. 20150013.

Henrich, Joseph and Francisco J. Gil-White (2001). "The Evolution of Prestige: Freely Conferred Deference as a Mechanism for Enhancing the Benefits of Cultural Transmission". *Evolution and Human Behavior* 22.3, pp. 165–196.

Hill, Ryan and Carolyn Stein (2025a). "Race to the Bottom: Competition and Quality in Science". *Quarterly Journal of Economics* 140.2, pp. 1111–1185. DOI: 10.1093/qje/qjaf010.

— (2025b). "Scooped! Estimating Rewards for Priority in Science". *Journal of Political Economy* 133.3, pp. 793–845. DOI: 10.1086/733398.

Hill, Ryan et al. (2025). "The Pivot Penalty in Research". *Nature* 642, pp. 999–1006. DOI: 10.1038/s41586-025-09048-1.

Huber, Jürgen, Michael Razen, Jörg Oechssler, et al. (2022). "Nobel and Novice: Author Prominence Affects Peer Review". *Proceedings of the National Academy of Sciences* 119.41, e2205779119.

Jiménez, Ángel V. and Alex Mesoudi (2019). "Prestige-Biased Social Learning: Current Evidence and Outstanding Questions". *Palgrave Communications* 5.1.

Jones, Benjamin F. (2009). "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?" *Review of Economic Studies* 76.1, pp. 283–317.

— (2010). "Age and Great Invention". *Review of Economics and Statistics* 92.1, pp. 1–14.

Jones, Benjamin F. and Bruce A. Weinberg (2011). "Age Dynamics in Scientific Creativity". *Proceedings of the National Academy of Sciences* 108.47, pp. 18910–18914.

Klein, Daniel B., ed. (1997). *Reputation: Studies in the Voluntary Elicitation of Good Conduct*. Economics, Cognition, and Society. Ann Arbor: University of Michigan Press.

Kreps, David M. and Robert Wilson (1982). "Reputation and Imperfect Information". *Journal of Economic Theory* 27.2, pp. 253–279.

Kuhn, Thomas S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Leamer, Edward E. (1983). "Let's Take the Con out of Econometrics". *American Economic Review* 73.1, pp. 31–43.

Lin, Zihang et al. (2023). "SciSciNet: A Large-Scale Open Data Lake for the Science of Science Research". *Scientific Data* 10.1, p. 315.

Liu, Lu et al. (2023). "Data, Measurement and Empirical Methods in the Science of Science". *Nature Human Behaviour* 7.7, pp. 1046–1058.

Merton, Robert K. (1957). "Priorities in Scientific Discovery: A Chapter in the Sociology of Science". *American Sociological Review* 22.6, pp. 635–659.

— (1961). "Singletons and Multiples in Scientific Discovery: A Chapter in the Sociology of Science". *Proceedings of the American Philosophical Society* 105.5, pp. 470–486.

— (1968). "The Matthew Effect in Science". *Science* 159.3810, pp. 56–63. DOI: 10.1126/science.159.3810.56.

Mokyr, Joel (2016). *A Culture of Growth: The Origins of the Modern Economy*. Princeton, NJ: Princeton University Press.

Myers, Kyle (2020). "The Elasticity of Science". *American Economic Journal: Applied Economics* 12.4, pp. 103–134.

Nelson, Phillip (1970). "Information and Consumer Behavior". *Journal of Political Economy* 78.2, pp. 311–329.

Nelson, Richard R. (1959). "The Simple Economics of Basic Scientific Research". *Journal of Political Economy* 67.3, pp. 297–306. DOI: 10.1086/258177.

Polanyi, Michael (1962). "The Republic of Science". *Minerva* 1.1, pp. 54–73. DOI: 10.1007/BF01101453.

Priem, Jason, Heather Piwowar, and Richard Orr (2022). *OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts*. Submitted to the 26th International Conference on Science, Technology and Innovation Indicators (STI 2022). DOI: 10.48550/arXiv.2205.01833. arXiv: 2205.01833 [cs.DL]. URL: https://arxiv.org/abs/2205.01833.

Schneider, Michael (2007). "The Nature, History and Significance of the Concept of Positional Goods". *History of Economics Review* 45.1, pp. 60–81.

Shapiro, Carl (1983). "Premiums for High Quality Products as Returns to Reputations". *Quarterly Journal of Economics* 98.4, pp. 659–679.

Sinha, Arnab et al. (2015). "An Overview of Microsoft Academic Service (MAS) and Applications". In: *Proceedings of the 24th International Conference on World Wide Web Companion*. Florence, Italy: Association for Computing Machinery. pp. 243–246. DOI: 10.1145/2740908.2742839.

Smith, Adam (1759). *The Theory of Moral Sentiments*. Ed. by D. D. Raphael and A. L. Macfie. Liberty Fund edition. Indianapolis: Liberty Fund, 1982.

Staiger, Douglas and James H. Stock (1997). "Instrumental Variables Regression with Weak Instruments". *Econometrica* 65.3, pp. 557–586. DOI: 10.2307/2171753.

Stephan, Paula E. (1996). "The Economics of Science". *Journal of Economic Literature* 34.3, pp. 1199–1235.

Stern, Scott (2004). "Do Scientists Pay to Be Scientists?" *Management Science* 50.6, pp. 835–853.

Sun, Liyang and Sarah Abraham (2021). "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects". *Journal of Econometrics* 225.2, pp. 175–199.

Tripodi, Giorgio et al. (2025). "Tenure and Research Trajectories". *Proceedings of the National Academy of Sciences* 122.30, e2500322122. DOI: 10.1073/pnas.2500322122.

Tullock, Gordon (1966). *The Organization of Inquiry*. Durham, NC: Duke University Press.

Wang, Kuansan et al. (Dec. 2019). "A Review of Microsoft Academic Services for Science of Science Studies". *Frontiers in Big Data* 2, p. 45. DOI: 10.3389/fdata.2019.00045.

Wang, Kuansan et al. (2020). "Microsoft Academic Graph: When Experts Are Not Enough". *Quantitative Science Studies* 1.1, pp. 396–413. DOI: 10.1162/qss{_}a{_}00021.

Wu, Lingfei, Dashun Wang, and James A. Evans (2019). "Large Teams Develop and Small Teams Disrupt Science and Technology". *Nature* 566, pp. 378–382. DOI: 10.1038/s41586-019-0941-9.

Appendix A: Proof of Inequality 5

Let reputations $\hat{\alpha} \in [0,1]$ have baseline CDF G. Let the thresholds satisfy $\alpha_A^* < \alpha_B^* < 1$. Define the active-seller CDF as the *conditional* CDF of reputations given survival above the threshold: for market i,

$$G_i(x) \equiv \Pr(\hat{\alpha} \le x \mid \hat{\alpha} \ge \alpha_i^*).$$

Thus, for $x \in [\alpha_i^*, 1)$,

$$G_i(x) = \frac{\Pr(\alpha_i^* \le \hat{\alpha} \le x)}{\Pr(\hat{\alpha} \ge \alpha_i^*)} = \frac{G(x) - G(\alpha_i^*)}{1 - G(\alpha_i^*)}.$$ [11]

For market $i \in \{A, B\}$, the active-seller CDF is

$$G_i(x) = \begin{cases} 0, & x < \alpha_i^*, \\ \dfrac{G(x) - G(\alpha_i^*)}{1 - G(\alpha_i^*)}, & \alpha_i^* \le x < 1, \\ 1, & x \ge 1. \end{cases} \tag{A1}$$

Claim. For all x, $G_B(x) \le G_A(x)$; the inequality is strict for $x \in [\alpha_B^*, 1)$ whenever $G(x) < 1$.

Proof. (i) If $x < \alpha_B^*$, then by (A1) $G_B(x) = 0 \le G_A(x)$.

(ii) If $\alpha_B^* \le x < 1$, set $a := G(x)$ and $b_i := G(\alpha_i^*)$. By monotonicity of G and $\alpha_A^* < \alpha_B^*$, we have $b_A \le b_B$. Define $f(b) := \frac{a-b}{1-b}$ for $b \in [0,1)$. Because $a < 1$, $f'(b) = \frac{a-1}{(1-b)^2} < 0$, so f is strictly decreasing. Thus

$$G_B(x) = f(b_B) \le f(b_A) = G_A(x),$$

with strict inequality if either $b_A < b_B$ or $a < 1$.

(iii) If $x \ge 1$, then $G_B(x) = G_A(x) = 1$.

Combining (i)–(iii) yields $G_B(x) \le G_A(x)$ for all x, with strict inequality on $[\alpha_B^*, 1)$ when $G(x) < 1$. Hence the active reputation distribution in B first-order stochastically dominates that in A. □

[11] If there is mass at α_i^* and "active" means $\hat{\alpha} \ge \alpha_i^*$, replace $G(\alpha_i^*)$ by the left limit $G(\alpha_i^{*-})$.

Appendix B: Regression analysis supplements

Table B1: Summary statistics for regression analysis

Panel A: Sample structure

Field–year	Mean	Median	Min	Max	p90	p99	N
Years	—	—	1900	2015	—	—	116
# Papers	29921	5235	96	856259	79196	337625	2204
# Authors	61213	6341	101	2173377	149684	831811	2204
# Titles used for *Testability*	2992	526	5	85658	7935	33700	2204

Subfield–year	Mean	Median	Min	Max	p90	p99	N
Years	—	—	1900	2015	—	—	116
# Papers	2289.	286	1	212064	5743.	29962.	26415
# Authors	5944.	364	1	745382	12402.	101774.	26415
# Titles used for *Testability*	146.	15	0	21180	351	2126.	54099

Panel B: Descriptive statistics

	Field–year				*Subfield–year*			
Variable	Mean	SD	p10	p90	Mean	SD	p10	p90
Testability (pctile 0–100)	50	28.71	11.3	89.98	50	23.4337	34.9335	90.0686
Paper-citations HHI	0.028	0.064	0.00024	0.0726	0.1419	0.2198	0.001433	0.426
— log	-5.28	2.155	-8.339	-2.622	-3.7768	3.8153	-6.5477	-0.8532
Author-citations HHI	0.023	0.054	0.00011	0.0573	0.1199	0.2037	0.00058	0.3543
— log	-5.661	2.351	-9.116	-2.859	-4.2026	3.918	-7.4526	-1.0376
Author h-index HHI	0.00474	0.0091	0.000027	0.0136	0.0779	0.1564	0.0002513	0.2175
— log	-7.117	2.346	-10.51	-4.299	-4.6552	3.0968	-8.289	-1.5255
Team size (median)	1.305	0.644	1	2	1.3797	0.7833	1	2
# Active authors	61213	175091	679	149684	5943.88	22940.45	21	12402.4

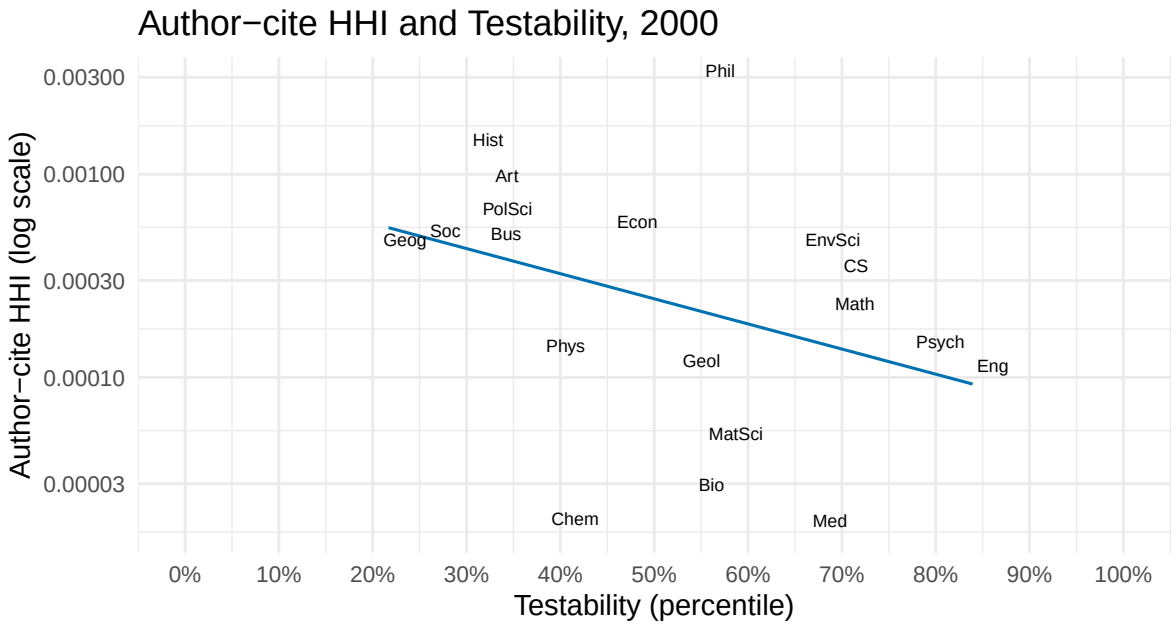


Figure B1: Sample year scatter plot of author citation HHI and testability.

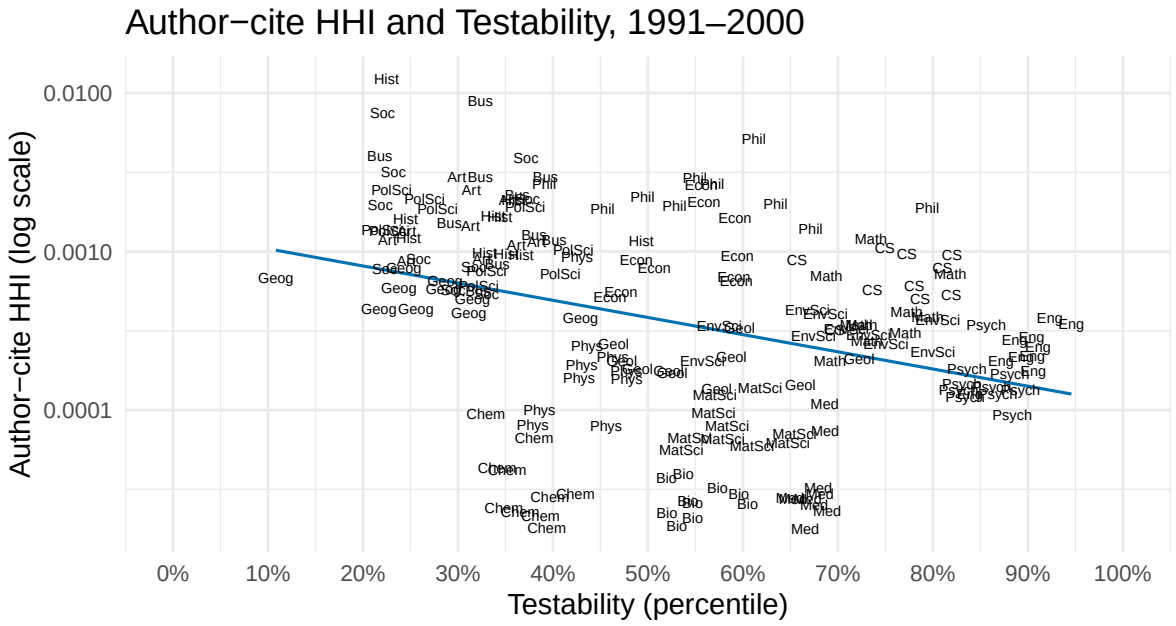


Figure B2: Sample decade scatter plot of author citation HHI and testability. Note that all regressions reported in Table 2 include year fixed effects.

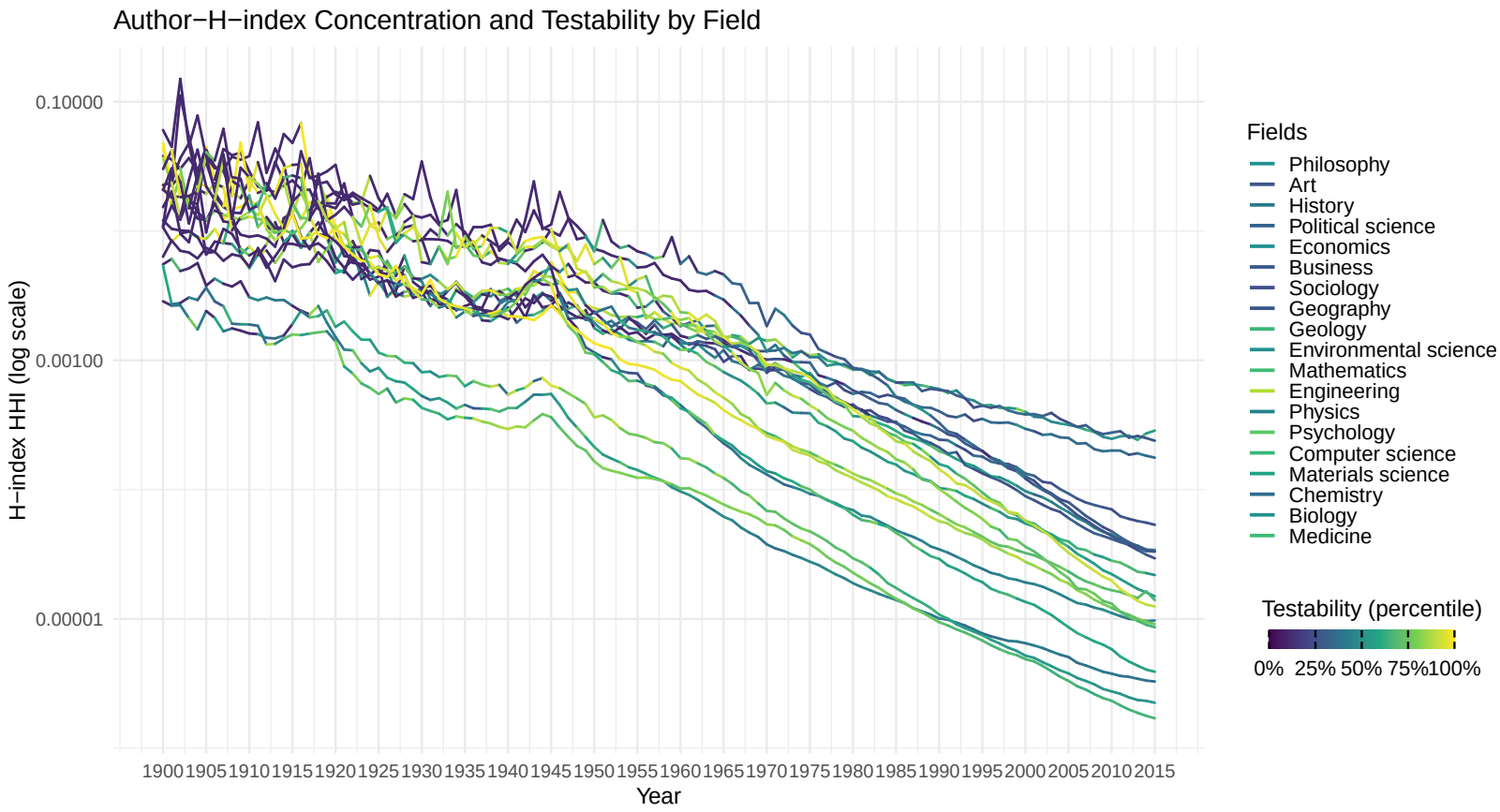


Figure B3: Author-H-index and Testability. *Note:* The prestige-testability tradeoff hypothesis says that there should be lower testability (dark) where there is higher concentration of prestige markers (top), but note that all regressions reported in Table 2 include year fixed effects. Testability is measured as the incidence of the string " test" (space included) in titles among the fields, and factored as percentile. HHI is calculated as $hhi_{f,t} = \sum_{i=1}^{n} \left(\frac{h-index_{i,t}}{h-index_{f,t}} \right)^2$, where $h-index$ is the lifetime h-index of an individual author i who is active in field f in year t; n is total active authors in that field-year. Number of field-years = 2204.The order of fields in the key is ranking in last observed year (2015).

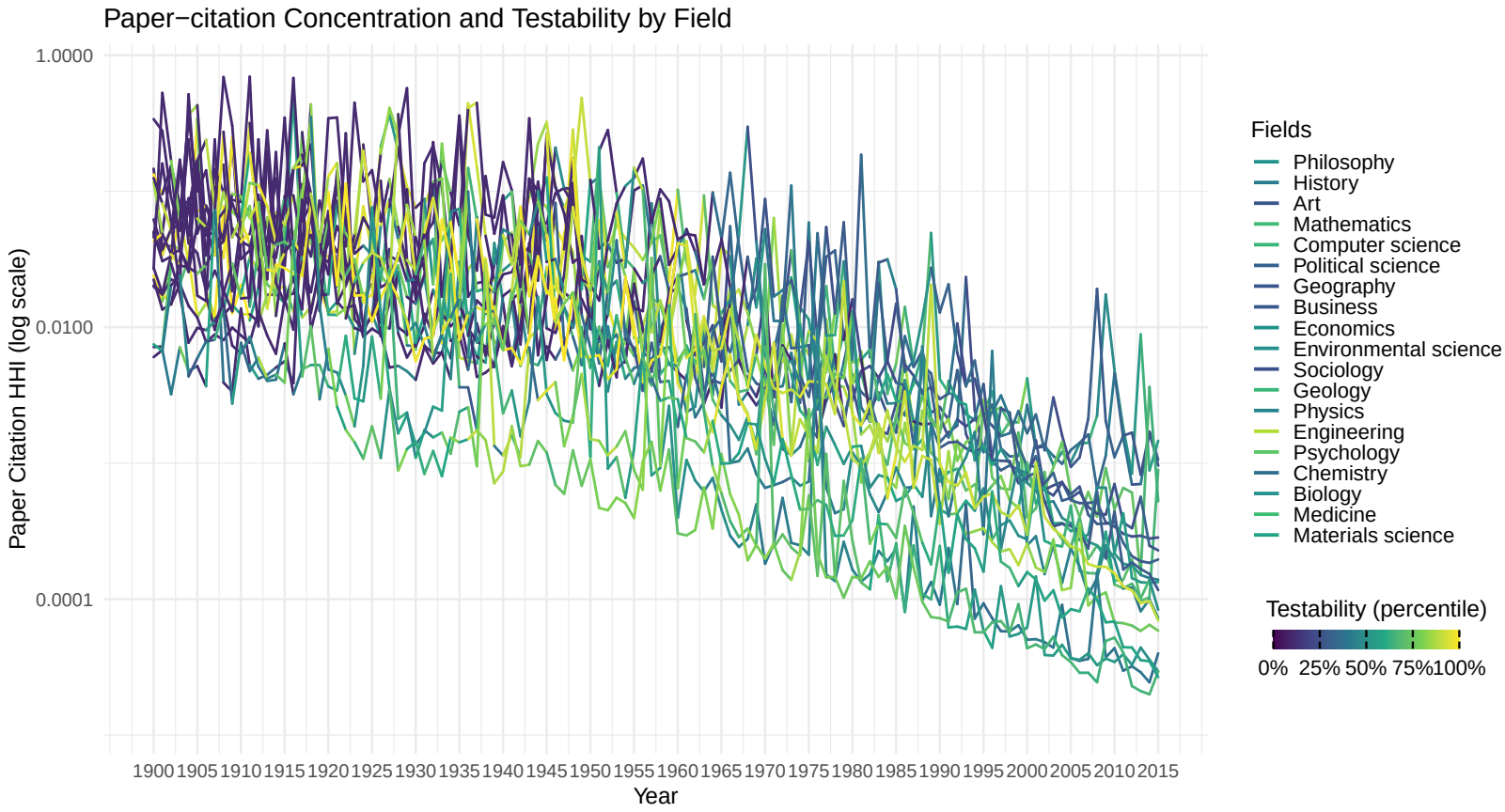


Figure B4: Paper-citation Concentration and Testability. *Note:* The prestige-testability tradeoff hypothesis says that there should be lower testability (dark) where there is higher concentration of prestige markers (top), but note that all regressions reported in Table 2 include year fixed effects. Testability is measured as the incidence of the string " test" (space included) in titles among the fields, and factored as percentile. HHI is calculated as $hhi_{f,t} = \sum_{i=1}^{n} \left(\frac{cites_{i,t}}{totalcites_{f,t}} \right)^2$, where f is a field, t is a year, i is an individual contribution, and n is number of contributions in the field. Number of field-years = 2204.The order of fields in the key is ranking in last observed year (2015).

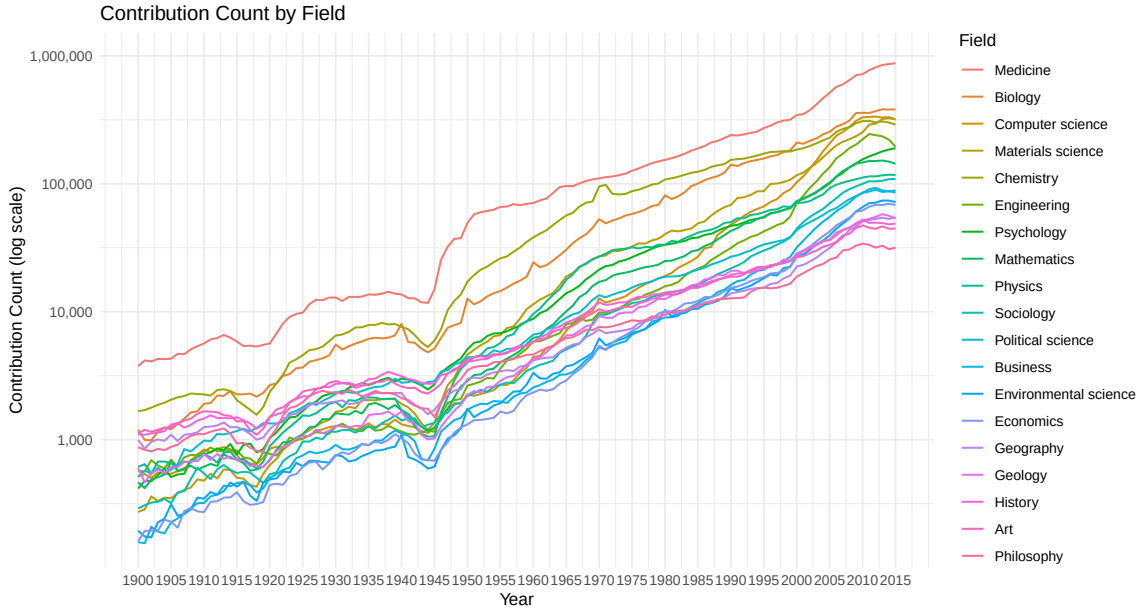


Figure B5: Contribution Count by Field (log scale). *Each contribution is counted as a unique Paper ID in SciSciNet. Here I include contributions identified as journal articles, books, book chapters, and conference papers, and I exclude those identified as datasets, thesis papers, repository papers (such as in ArXive or SSRN) or are left unidentified. This yields 91,479,382 out of the approximately 134 million total contributions in SciSciNet*

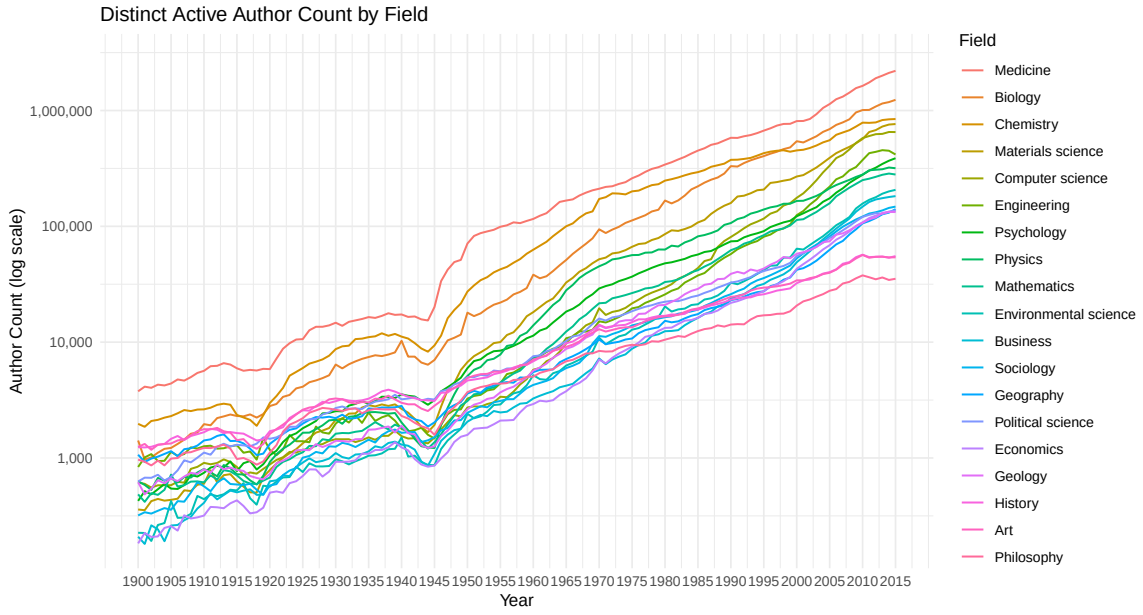
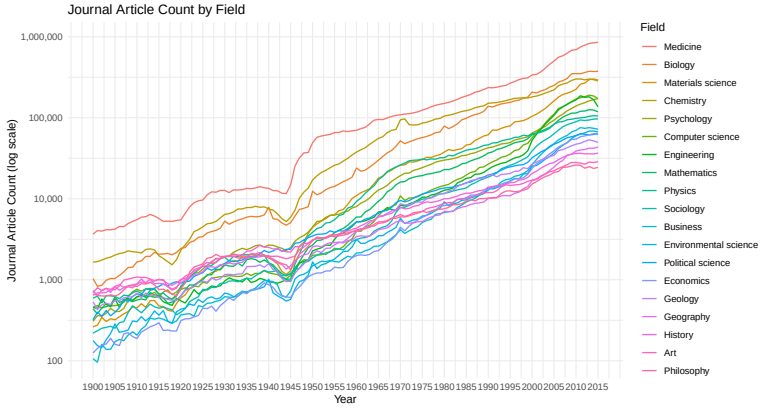
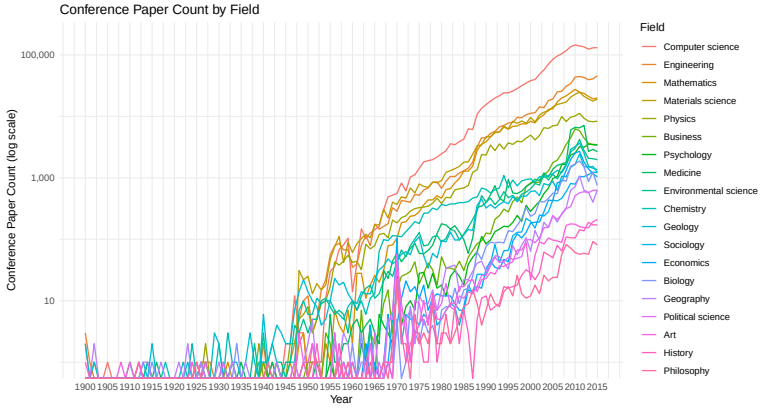


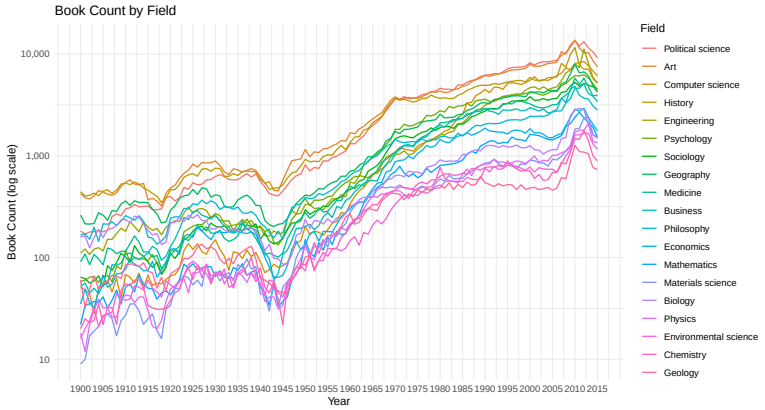
Figure B6: Author Count by Field (log scale). *Each author is counted as a unique Author ID in SciSciNet with at least one contribution in a field-year. See Lin et al. 2023, 5, for the method used for author disambiguation.*



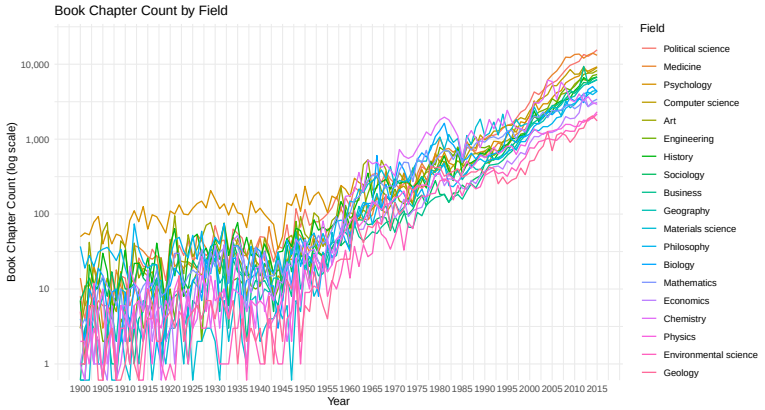
(a) Journal articles. *74,013,927 of ≈ 134M.*



(b) Conference papers. *4,874,808 of ≈ 134M.*



(c) Books. *3,121,693 of ≈ 134M.*



(d) Book chapters. *2,230,448 of ≈ 134M.*

Figure B7: Contribution Count by Type (log scale): journal articles, conference papers, books, and book chapters.

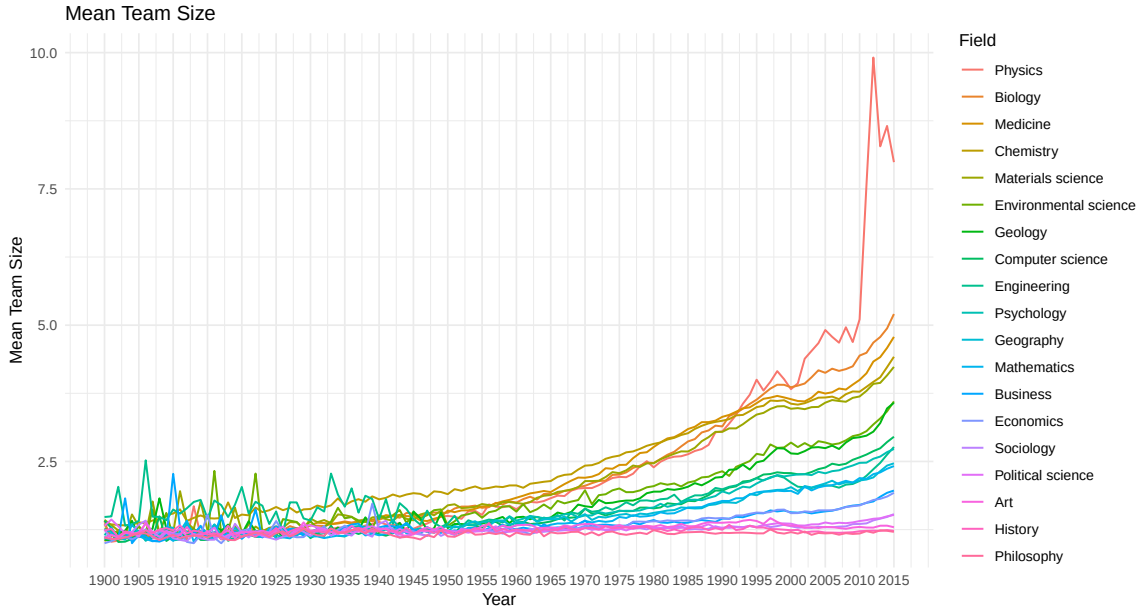


Figure B8: Mean Team Size by Field. *Each team size is calculated as the number of distinct authors on a contribution. These are generated from a 10% sample of contributions in each field-year.*

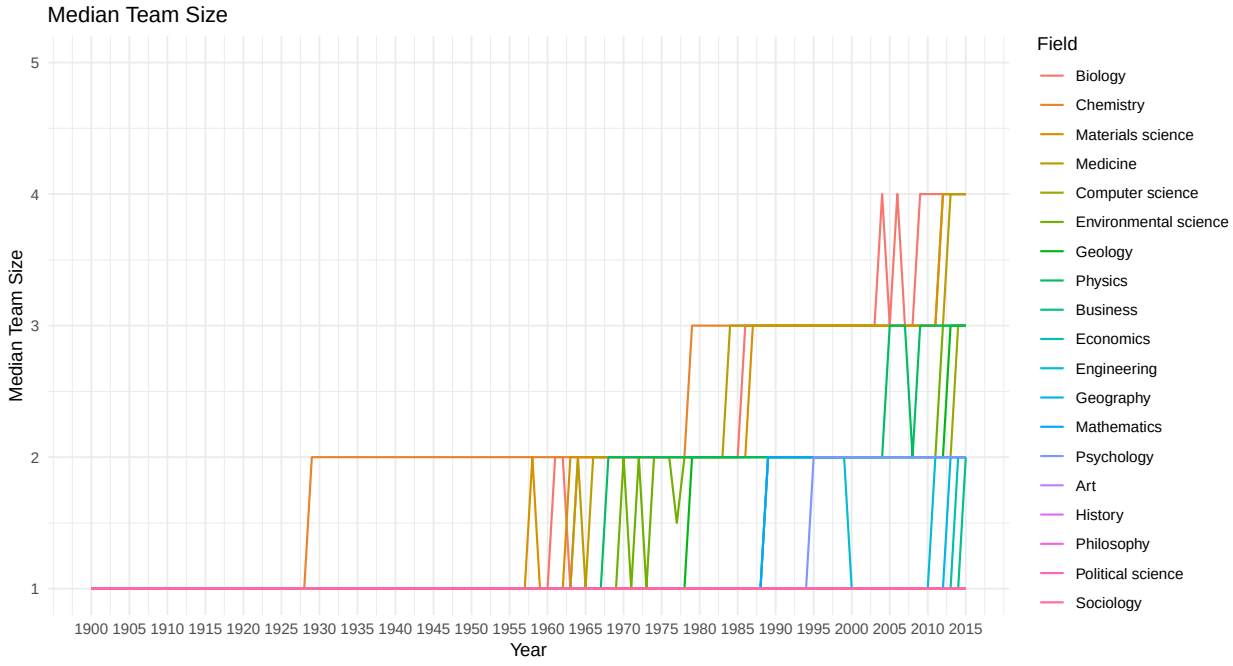


Figure B9: Median Team Size by Field. *Each team size is calculated as the number of distinct authors on a contribution. These are generated from a 10% sample of contributions in each field-year.*

Appendix C: Event study supplements

C.1 Interpreting the Sun–Abraham Event-Study Estimates with Potential Outcomes.

An author i joins cohort g when she cites a seed-paper in her adoption year $C_i \in \mathcal{G} \cup \{\infty\}$; never-adopters have adoption year $C_i = \infty$. Event time is defined as $k = t - C_i$, where t is calender time. The fact that authors' adoption takes place over calender time give the impetus for the staggered event study design. Potential outcomes are $Y_{it}(c)$, the outcome for author i in year t if i were to adopt in year c (with $c = \infty$ the never-adopt state).

For cohort g and event time k, the causal effect of interest is

$$\tau_g(k) \;=\; \mathbb{E}[\, Y_{i,\, g+k}(g) \;-\; Y_{i,\, g+k}(\infty) \mid C_i = g \,].$$

That is, given that some author is in cohort g, the difference between their realized outcome and their unrealized potential outcome of being untreated is the causal effect to which I approximate with the following estimation and assumptions.

I estimate an interaction-weighted event study as presented in Sun and Abraham (2021) with reference period $k = -1$ omitted and fixed effects μ that are either calendar-year FE (Main) or subfield-by-year FE (Strict). Let $\widehat{\beta}_k$ denote the estimated coefficient at event time k. Under the assumptions below,

$$\widehat{\beta}_k \;\approx\; \sum_{g \in \mathcal{G}_k} w_{gk}\, \tau_g(k), \qquad \sum_{g \in \mathcal{G}_k} w_{gk} = 1,$$

that is, a cohort-weighted average of cohort-specific causal effects for the cohorts $\mathcal{G}_k$ that contribute observations at event time k. The weights w_{gk} reflect the Sun–Abraham interaction-weighting across cohorts.

Counterfactual and Comparison Set. Conceptually, $\tau_g(k)$ compares the path under adoption in the observed cohort g to the *never-adopt* path $Y_{it}(\infty)$. Empirically, at each calendar year t, the control group used by the estimator consists of *not-yet-treated* (and any never-treated) authors, $\{i : C_i > t\}$; already-treated authors ($C_i \leq t$) do not serve as controls.

Using authors who are not-yet-treated as a comparison group comes with advantages and disad-

vantages. Firstly, if there are hidden variables that cause assortment into *ever*-treatment status, this technique can count that out as a determinant of any estimated causal effects; I take this fact as a strong argument for using the present design, given that authors surely non-randomly choose fields, topics, methods, and ultimately treatment status. However, by construction the comparison group is always part of a group that adopts later in calender time; this would introduce bias if calender time of treatment correlates with effect sizes. For this reason, I include a calender year fixed effect in all my estimates.

Heterogeneity by Pre-Treatment Outcome. To study whether effects differ across researchers with different *pre-treatment* outcome levels, I partition authors into G groups using only information from the period before adoption. For each outcome Y (e.g., $c5$, $c10$, hit10, or top5), I construct a baseline statistic

$$S_i(Y) \;=\; \text{median}_{t<C_i}\; Y_{it},$$

the author-specific median outcome over pre-adoption years.[12] I then rank authors by $S_i(Y)$ and assign a group label $Q_i \in \{1, \ldots, G\}$ (e.g., $G = 2$ for a median split, $G = 3$ for terciles, or as I present $G = 5$ for quantiles). Importantly, the grouping uses only pre-treatment data, so it is unaffected by post-adoption dynamics.

Given these groups, I estimate an interacted Sun–Abraham specification that allows the dynamic treatment path to differ across Q_i:

$$Y_{it} = \sum_{k\neq -1} \sum_{q=1}^{G} \beta_k^{(q)} \left(\sum_{g\in\mathcal{G}} \mathbf{1}\{C_i = g\}\, \mathbf{1}\{t - g = k\} \right) \mathbf{1}\{Q_i = q\} \;+\; \mu \;+\; \varepsilon_{it},$$

with the same fixed effects μ as in the main specification (calendar-year FE for Main; subfield×year FE for Strict) and the same weighting choice (unweighted or w_{it}). Let $\widehat{\beta}_k^{(q)}$ denote the coefficient for group q at event time k (normalized so $\widehat{\beta}_{-1}^{(q)} = 0$). Under the standard identification conditions,

$$\widehat{\beta}_k^{(q)} \;\approx\; \sum_{g\in\mathcal{G}_k} w_{gk}^{(q)}\, \mathbb{E}\Big[\, Y_{i,\,g+k}(g) - Y_{i,\,g+k}(\infty) \;\Big|\; C_i = g,\; Q_i = q \Big], \qquad \sum_{g\in\mathcal{G}_k} w_{gk}^{(q)} = 1,$$

so each $\widehat{\beta}_k^{(q)}$ can be read as the average (across contributing cohorts) causal effect at event time k for authors who started in heterogeneity group q, relative to their never-adopter counterfactual, with

[12] For never-adopters ($C_i = \infty$), the baseline $S_i(Y)$ is computed over all observed years. Authors without any pre-period observations for a given outcome are not assigned to a group for that outcome. Ties are broken deterministically.

$k = -1$ as the pre-adoption reference.

Interpretation. The group-specific paths $\{\widehat{\beta}_k^{(q)}\}$ trace out how treatment effects evolve for authors with "lower" vs. "higher" baselines. Differences across groups reflect systematic heterogeneity present *before* adoption, not changes induced by treatment. This focus on pre-treatment heterogeneity is integral for my interpretation of a prestige-testability tradeoff. As with the main estimates, valid interpretation requires (i) no anticipation; (ii) parallel trends in the never-adopt state conditional on the fixed effects; and (iii) adequate support from not-yet-treated controls within the fixed-effect cells for each group and event time.

Assumptions for Causal Interpretation.

1. **No anticipation.** For any cohort g,

$$Y_{it}(g) \;=\; Y_{it}(\infty) \quad \text{for all } t < g.$$

2. **Parallel trends in the never-adopt path (conditional on FE).** For any cohorts g_1, g_2 and any $t < \min\{g_1, g_2\}$,

$$\mathbb{E}\big[\, Y_{i,t+\Delta}(\infty) - Y_{it}(\infty) \;\big|\; C_i = g_1, \mathrm{FE}\big] \;=\; \mathbb{E}\big[\, Y_{i,t+\Delta}(\infty) - Y_{it}(\infty) \;\big|\; C_i = g_2, \mathrm{FE}\big]$$

   for all feasible $\Delta \geq 0$ with $t + \Delta < \min\{g_1, g_2\}$. (In the heterogeneity runs, the same condition holds *within* each pre-treatment group $Q_i = q$.)

3. **SUTVA / no interference.** Let D_{-i} denote others' adoption histories. Then

$$Y_{it}(c, D_{-i}) \;=\; Y_{it}(c) \quad \text{for all } c \in \mathcal{G} \cup \{\infty\},\; t,\; D_{-i}.$$

   (Outcomes for author i are unaffected by other authors' adoption timing.)

4. **Heterogeneous treatment effects allowed; group-specific parallel trends.** Cohort- and time-specific effects $\tau_g(k)$ (and, with heterogeneity, $\tau_g^{(q)}(k)$) may vary arbitrarily. Identification relies on the parallel-trends condition within $Q_i = q$; no homogeneity is assumed. Formally, for

any g_1, g_2 and all $t, s < \min\{g_1, g_2\}$,

$$\begin{aligned} &\mathbb{E}\big[\, Y_{it}(\infty) \;\big|\; C_i = g_1, Q_i = q, \mathrm{FE}\big] - \mathbb{E}\big[\, Y_{is}(\infty) \;\big|\; C_i = g_1, Q_i = q, \mathrm{FE}\big] \\ =\; &\mathbb{E}\big[\, Y_{it}(\infty) \;\big|\; C_i = g_2, Q_i = q, \mathrm{FE}\big] - \mathbb{E}\big[\, Y_{is}(\infty) \;\big|\; C_i = g_2, Q_i = q, \mathrm{FE}\big]\,. \end{aligned}$$

Discussion of assumptions and interpretation.

1. **No anticipation.** Authors do not adjust outcomes *before* their first seed-citation year because of that future adoption. However, in reality, authors "truly adopt" before the final publication, as it requires "time to build" a publication, often over multiple years. In practice, anticipatory behavior that effects outcomes would show changes before the adoption date. For transparency, I present the main results without accounting for anticipatory behavior.

2. **Parallel trends (never-adopt path).** Conditional on the fixed effects, cohorts would have followed the same evolution in the counterfactual "never-adopt" state. Year FE (Main) soak up aggregate time shocks; subfield×year FE (Strict) absorb field-specific time shocks. Remaining differential drift across cohorts within those cells would bias the SA contrasts; the pre-trend tests are designed to detect this, but, as always, parallel trends cannot be tested directly.

3. **SUTVA / Spillovers.** One author's adoption should not mechanically alter another author's outcome except through common shocks already absorbed by FE. If spillovers are plausible (e.g. peers' adoption affects citations), this would bias estimates. Because I cannot rule this out, we may discount the precision of the estimates as follows.

   Let $Y_{it}(d, s)$ denote author i's potential outcome in year t when her own treatment status is $d \in \{0, 1\}$ (untreated/treated by t) and her exposure to others' adoption is $s \in [0, 1]$ (e.g., the share of peers/coauthors/field colleagues already treated at t). SUTVA corresponds to $Y_{it}(d, s) = Y_{it}(d, 0)$ for all s (no exposure effect). Our target at event time k for cohort g is

$$\tau_g(k) \;=\; \mathbb{E}[\, Y_{i,\, g+k}(1, 0) \;-\; Y_{i,\, g+k}(0, 0) \,|\, C_i = g\,]\,.$$

   With spillovers, the IW estimator contrasts treated observations with not-yet-treated controls

who may have $s > 0$, yielding

$$\widehat{\beta}_k \;\approx\; \tau_g(k) \;+\; B_{gk}, \qquad B_{gk} \;=\; \underbrace{\mathbb{E}\big[Y_{g+k}(1, S^T) - Y_{g+k}(1,0)\big]}_{\text{spillover on treated}} \;-\; \underbrace{\mathbb{E}\big[Y_{g+k}(0, S^C) - Y_{g+k}(0,0)\big]}_{\text{spillover on controls}},$$

where S^T and S^C are the (possibly different) exposure levels faced by treated and control groups at (g, k).

For sign intuition, adopt a simple additive exposure model

$$Y_{it}(d, s) \;=\; Y_{it}(d, 0) \;+\; \theta_t^{(d)}\, s,$$

with $\theta_t^{(1)}$ (spillover slope when treated) and $\theta_t^{(0)}$ (when untreated). Then

$$B_{gk} \;=\; \theta_{g+k}^{(1)}\, \mathbb{E}[S^T] \;-\; \theta_{g+k}^{(0)}\, \mathbb{E}[S^C].$$

- **Case 1 (positive spillover onto controls ⇒ underestimate).**

  Example: As colleagues begin citing the seed methods, overall attention to the topic rises, lifting citations even for not-yet-treated authors, so the control group improves too.

  Or: If $\theta_{g+k}^{(0)} > 0$ and $\mathbb{E}[S^C] > 0$ (controls benefit), while $\theta_{g+k}^{(1)}\,[S^T]$ is comparable or smaller, then

  $$B_{gk} \;\approx\; -\,\theta_{g+k}^{(0)}\, \mathbb{E}[S^C] \;<\; 0,$$

  so $\widehat{\beta}_k$ is biased *toward zero*: it *underestimates* $\tau_g(k)$.

- **Case 2 (negative spillover onto controls ⇒ overestimate).**

  Example: Adoption pulls attention away from not-yet-treated authors (e.g., referees or readers penalize older approaches), depressing their outcomes and widening the treated–control gap.

  Or: If $\theta_{g+k}^{(0)} < 0$ and $\mathbb{E}[S^C] > 0$ (controls are hurt), then

  $$B_{gk} \;\approx\; -\,\theta_{g+k}^{(0)}\, \mathbb{E}[S^C] \;>\; 0,$$

  so $\widehat{\beta}_k$ is biased *upward*: it *overestimates* $\tau_g(k)$.

At this point, I cannot yet determine if these effects are substantial or which one potentially

dominates, so estimates should be discounted accordingly. Note: Field–year fixed effects absorb shocks common to all authors in a subfield–year, but they do not remove differential exposure within that cell.

4. **Heterogeneity.** When I split authors by pre-treatment heterogeneity group Q_i of a given outcome (computed only from pre-adoption years), the SA paths are estimated separately within each group. Interpretation of $\widehat{\beta}_k^{(q)}$ requires the same parallel-trends logic to hold *within* group q. Differences across groups then reflect genuine effect heterogeneity present already before adoption, rather than sorting on post-treatment dynamics.

5. **Normalization and Fixed Effects.** Coefficients are normalized so that $\widehat{\beta}_{-1} = 0$. The Main specification includes year FE (μ_t), while the Strict specification includes subfield-by-year FE ($\mu_{\ell t}$), absorbing common shocks at the corresponding aggregation.

C.2 Event study summary statistics

Table C1: Summary statistics for event study analysis

Panel A: Sample structure

Main: Year–FE	Mean	Median	Min	Max	p90	p99	N
Years	—	—	1993	2015	—	—	23
Authors per year	1004.870	1048.000	441	1,470	1,450	1,470	23
Treated share per year	0.359	0.298	0.005	1.000	0.773	0.975	23
Strict: Subfield × Year FE	Mean	Median	Min	Max	p90	p99	N
Years	—	—	1993	2015	—	—	23
Authors per subfield–year	14.697	6.000	1	210	37	134	1,411
Treated share per subfield–year	0.596	0.571	0.000	1.000	1.000	1.000	1,411

Panel B: Descriptive statistics

	Year–FE				*Subfield × Year FE*			
Variable	Mean	SD	p10	p90	Mean	SD	p10	p90
Citations (z, c5)	0.549	0.868	-0.234	1.733	0.565	0.874	-0.225	1.759
Citations (z, c10)	0.474	0.836	-0.245	1.613	0.508	0.857	-0.234	1.680
Hit (10y)	0.338	0.375	0.000	1.000	0.345	0.378	0.000	1.000
Top-5 share	0.036	0.158	0.000	0.000	0.037	0.160	0.000	0.000

Table C2: Treatment cohorts by adoption year — Year FE and Subfield × Year FE

	Year FE sample					*Subfield × Year FE sample*				
Adoption Cohort Year	Treated	Share (%)	Cum.	Cum. share (%)	Total	Treated	Share (%)	Cum.	Cum. share (%)	Total
1993	2	0.1	2	0.1	3,284	2	0.1	2	0.1	3,260
1994	5	0.2	7	0.2	3,284	4	0.1	6	0.2	3,260
1995	24	0.7	31	0.9	3,284	22	0.7	28	0.9	3,260
1996	25	0.8	56	1.7	3,284	25	0.8	53	1.6	3,260
1997	38	1.2	94	2.9	3,284	37	1.1	90	2.8	3,260
1998	45	1.4	139	4.2	3,284	44	1.3	134	4.1	3,260
1999	70	2.1	209	6.4	3,284	70	2.1	204	6.3	3,260
2000	73	2.2	282	8.6	3,284	73	2.2	277	8.5	3,260
2001	75	2.3	357	10.9	3,284	73	2.2	350	10.7	3,260
2002	76	2.3	433	13.2	3,284	74	2.3	424	13.0	3,260
2003	108	3.3	541	16.5	3,284	105	3.2	529	16.2	3,260
2004	119	3.6	660	20.1	3,284	119	3.7	648	19.9	3,260
2005	91	2.8	751	22.9	3,284	88	2.7	736	22.6	3,260
2006	156	4.8	907	27.6	3,284	156	4.8	892	27.4	3,260
2007	156	4.8	1,063	32.4	3,284	156	4.8	1,048	32.1	3,260
2008	187	5.7	1,250	38.1	3,284	187	5.7	1,235	37.9	3,260
2009	215	6.5	1,465	44.6	3,284	211	6.5	1,446	44.4	3,260
2010	215	6.5	1,680	51.2	3,284	215	6.6	1,661	51.0	3,260
2011	316	9.6	1,996	60.8	3,284	316	9.7	1,977	60.6	3,260
2012	273	8.3	2,269	69.1	3,284	272	8.3	2,249	69.0	3,260
2013	320	9.7	2,589	78.8	3,284	317	9.7	2,566	78.7	3,260
2014	332	10.1	2,921	88.9	3,284	331	10.2	2,897	88.9	3,260
2015	363	11.1	3,284	100.0	3,284	363	11.1	3,260	100.0	3,260

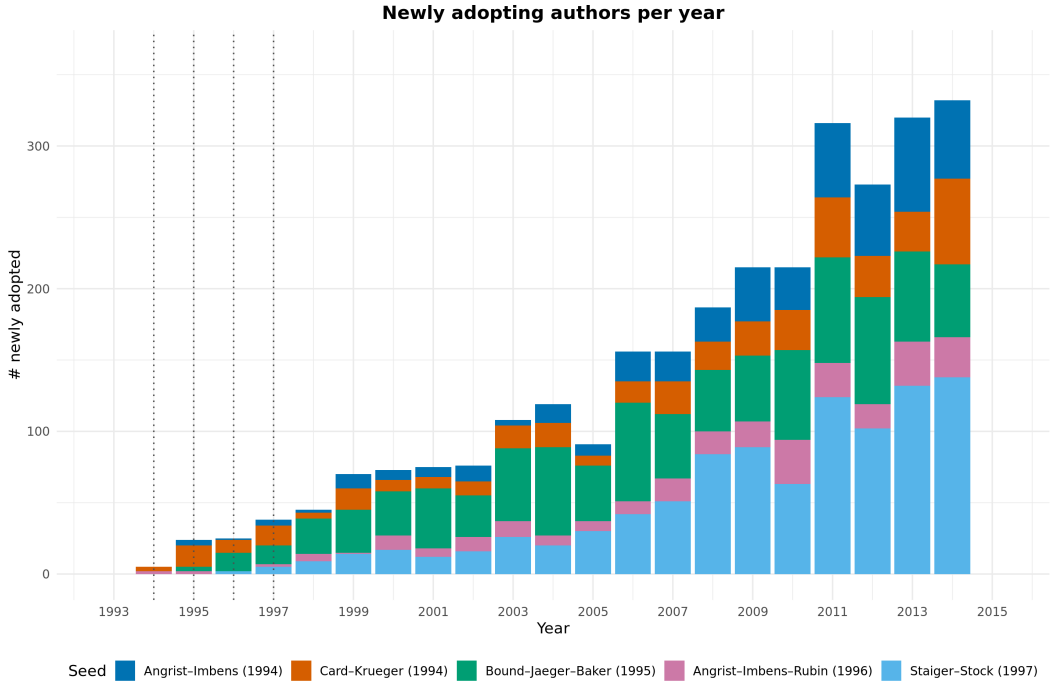


Figure C1: Newly adopting authors each year, where adoption is citing a seed paper. Each adoption year comprises a distinct cohort used in the staggered event study estimates. $N = 3,284$.

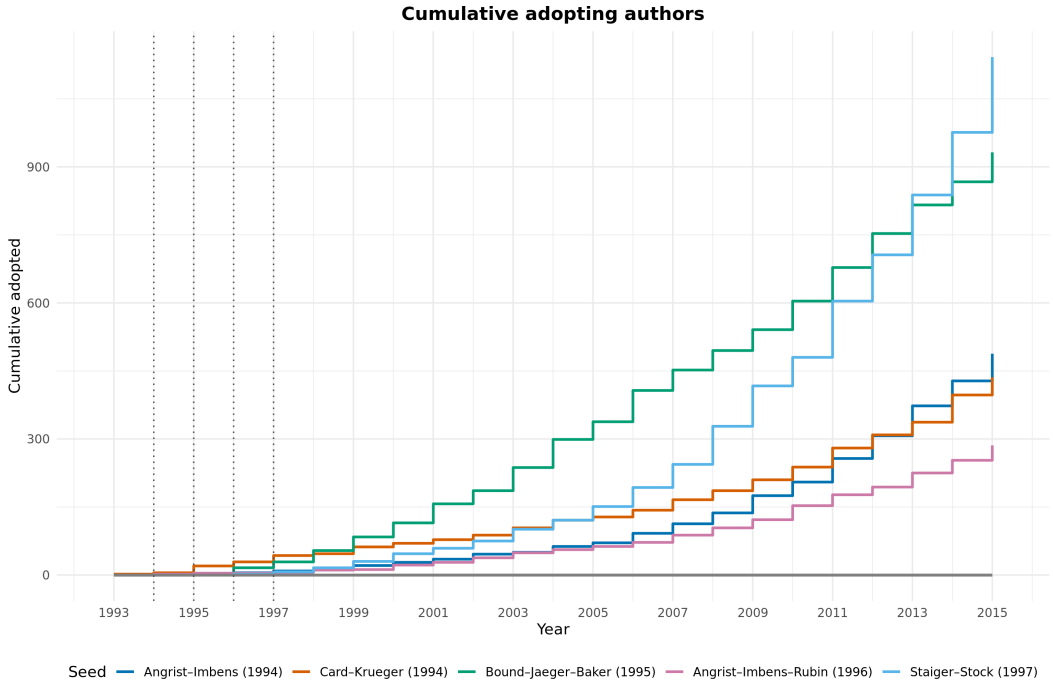


Figure C2: Cumulative adopters over time, where adoption is citing a seed paper. $N = 3,284$.

Table C3: Risk support (left) and estimation support (right) by event time k (NYT-only, main-seed authors)

Panel A: Risk (NYT, Year–FE)

k	Rows	Mass@k	Risk@(-1)	Cohorts	Min cohort	Max cohort
-5	7,292	7,292	1,516	18	1998	2015
-4	1,161	1,161	1,533	19	1997	2015
-3	1,274	1,274	1,552	20	1996	2015
-2	1,368	1,368	1,569	21	1995	2015
-1	1,572	1,572	1,572	22	1994	2015
0	2,262	2,262	1,572	23	1993	2015
1	1,299	1,299	1,408	21	1994	2014
2	1,117	1,117	1,249	21	1993	2013
3	955	955	1,094	20	1993	2012
4	869	869	953	19	1993	2011
5	689	689	800	18	1993	2010
6	601	601	702	17	1993	2009
7	484	484	597	16	1993	2008
8	420	420	503	15	1993	2007
9	344	344	428	13	1994	2006
10	1,405	1,405	364	12	1994	2005

Panel B: Risk (NYT, Subfield × Year FE)

k	Rows	Mass@k	Risk@(-1)	Cohorts	Min cohort	Max cohort
-5	5,362	5,362	1,451	18	1998	2015
-4	1,042	1,042	1,460	19	1997	2015
-3	1,173	1,173	1,470	20	1996	2015
-2	1,252	1,252	1,470	20	1996	2015
-1	1,476	1,476	1,476	21	1995	2015
0	2,258	2,258	1,476	23	1993	2015
1	1,294	1,294	1,315	21	1994	2014
2	1,117	1,117	1,161	21	1993	2013
3	954	954	1,009	20	1993	2012
4	869	869	871	19	1993	2011
5	688	688	718	18	1993	2010
6	601	601	622	17	1993	2009
7	484	484	519	16	1993	2008
8	419	419	428	15	1993	2007
9	344	344	358	13	1994	2006
10	1,404	1,404	295	12	1994	2005

Panel C: Estimation (NYT, Year–FE)

k	Contrib. cohorts	Mass@k	Risk@(-1)	Effective cohorts	Min cohort	Max cohort
-5	18	7,292	1,516	17.0	1998	2015
-4	19	1,161	1,533	13.7	1997	2015
-3	20	1,274	1,552	13.9	1996	2015
-2	21	1,368	1,569	14.3	1995	2015
-1	22	1,572	1,572	14.4	1994	2015
0	22	2,260	1,572	14.1	1994	2015
1	21	1,299	1,408	13.4	1994	2014
2	20	1,116	1,249	12.7	1994	2013
3	19	954	1,094	12.8	1994	2012
4	18	868	953	12.1	1994	2011
5	17	688	800	11.8	1994	2010
6	16	600	702	11.1	1994	2009
7	15	483	597	10.7	1994	2008
8	14	419	503	10.8	1994	2007
9	13	344	428	10.0	1994	2006
10	12	1,405	364	7.4	1994	2005

Panel D: Estimation (NYT, Subfield × Year FE)

k	Contrib. cohorts	Mass@k	Risk@(-1)	Effective cohorts	Min cohort	Max cohort
-5	18	5,362	1,451	15.1	1998	2015
-4	19	1,042	1,460	12.3	1997	2015
-3	20	1,173	1,470	12.9	1996	2015
-2	20	1,252	1,470	13.1	1996	2015
-1	21	1,476	1,476	13.4	1995	2015
0	21	2,253	1,476	14.1	1995	2015
1	20	1,291	1,315	13.4	1995	2014
2	19	1,113	1,161	12.6	1995	2013
3	18	950	1,009	12.7	1995	2012
4	17	866	871	12.1	1995	2011
5	16	684	718	11.7	1995	2010
6	15	596	622	11.0	1995	2009
7	14	480	519	10.6	1995	2008
8	13	415	428	10.6	1995	2007
9	12	341	358	9.8	1995	2006
10	11	1,401	295	7.4	1995	2005

Notes: **Risk panels (left):** *Rows* are author–year observations at event time k. *Mass@k* is the identification mass at k (controls for leads $k < 0$, treated for lags $k \geq 0$). *Risk@(-1)* is the number of control rows at $k = -1$ for the *same* cohorts that appear at k. *Cohorts* counts distinct cohort years at k (Min/Max are their bounds). **Estimation panels (right):** *Contrib. cohorts* appear at k and have nonzero control mass at $k = -1$; only these cohorts are used. *Effective cohorts* $= 1/\sum_g w_g^2$, with w_g proportional to the identification mass at k (controls for leads, treated for lags) and normalized to sum to one across contributing cohorts.