Online Appendix for:
Scooped! Estimating Rewards for Priority in Science

Ryan Hill Carolyn Stein

September 5, 2024

A Data Appendix

A.1 Protein Data Bank

The Protein Data Bank (PDB) is the main source of project data we use to construct priority races. The first iteration of the PDB started in 1971, and the current archive is a global collaboration run by a non-profit organization called the World Wide Protein Data Bank (wwPDB). The wwPDB is a union of four existing data banks from around the world, including the Research Collaboratory for Structural Bioinformatics Protein Database (RCSB PDB), Protein Data Bank in Europe (PDBe), Protein Data Bank Japan (PDBj), and Biological Magnetic Resonance Data Bank (BMRB). The data has been standardized and currently represents the universe of discoveries deposited in each of these archives. All new discoveries deposited to any database are transferred to, processed, standardized, and archived by the RCSB (Berman et al. 2006) at Rutgers University. Details about the PDB data can be found on their website.[1]

We access the data directly from the RCSB Custom Report Web Service.[2] The data extract used in this study was downloaded on May 22, 2018. We use the following field reports and variables:

- Structure Summary: structure ID, structure title, structure authors, deposit date, release date.
- Citation: PubMed ID, publication year, and journal name.
- Cluster Entity: entity ID, chain ID, sequence similarity clusters (BLAST algorithm for 90 percent and 100 percent sequence similarity, see section B below)
- Data Collection Details: collection date (the self-reported date the scientists generated diffraction data at a major synchrotron or in a home lab. Five percent of structures have multiple collections dates, so we keep the earliest.).

Additional data on cluster entities was accessed through a separate raw file archive at RCSB[3] on December 14, 2018. These files provided additional cluster groupings for the BLAST algorithm at 50 percent and 70 percent sequence similarity.

[1] http://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction
[2] https://www.rcsb.org/pdb/results/reportField.do
[3] ftp://resources.rcsb.org/sequence/clusters/ clusters50.txt and clusters70.txt

Supplemental Material (not copyedited or formatted) for: Ryan Hill, Carolyn Stein. 2025. "Scooped! Estimating Rewards for Priority in Science."
Journal of Political Economy 133(3). DOI: https://doi.org/10.1086/733398.

A.2 Citations and Journal Impact Factor

We use the journal names from the PDB extracts to link data to the Journal Citations Reports for journal impact factor and the Web of Science for citations.[4] We link the Journal Citations Reports using the journal name listed in the PDB. Each journal has an impact factor in each year and is calculated as the average number of citations per paper in the preceding two years. The JIF data was only avialable between 1997 and 2017, so we imputed impact factors in years before or after that window with the 1997 or 2017 impact factors respectively. We standardize impact factor in each year within the set of PDB-linked publications in our extracts each year. The citation data from the Web of Science and is restricted to citations from papers linked to PubMed IDs,[5] and self-citations are excluded. Citations are aggregated for each cited paper by publication year of the citing paper. When we report five-year citations, it represents the total number of citations in the publishing year and the subsequent five calendar years.

A.3 Altmetric.com Data

We use data from Altmetric.com to measure alternative forms of attention for academic research.[6] One limitation of the Altmetric data extract we use is that it only reports cumulative counts from the time of publication to the present (date of access: August 2nd, 2019). We account for the fact that scooped papers are published later and have less time to accumulate attention scores, using information about the change in score in recent time periods. The Altmetric.com data reports the change in attention in the past week, month, etc. We can therefore restrict the regression sample to races in which both teams had not accrued any additional attention in the amount of time that had passed between publications. For example, if paper A was released two months before paper B, we do not include this race in the analysis if paper A or paper B had accrued any additional attention in the most recent two months. This allows paper B to have the same window of time to accrue attention despite starting two months later. Because races in our sample end across a wide range of years, the regression coefficients are interpreted as the percent difference in outcomes for papers of an average vintage.

A.4 Editorial Dates

We access the received, accepted, and published dates from the websites of publications of Science, Nature Journals, Cell Press, and Public Library of Science. While we were not able to obtain these dates for all articles, we chose to focus on journals based on their prevalence in the PDB and the availability of the data for download. This subsample covers 19 percent of our primary regression sample.

We use these data to look at the correspondence between the journal publication date and the release date. Appendix Figure A4 reports the correspondence between the PDB release date and the publication date for the 616 articles in the racing sample for which they are available. This correspondence is not exact for a few reasons. First, according to PDB policy, scientists are allowed to release their findings immediately after deposit, which could potentially come before the publication date. In typical practice, the scientists prefer to wait until publication so that other scientists cannot use the information for follow-on work until after publication. In fact, scientists prefer to wait for release as long as possible to maintain a competitive advantage, which was the motivation behind the 1998 policy change to align release and

[4] Both data sources were owned by Thompson Reuters at the time of access, but have since been sold to Clarivate Analytics.

[5] Because structural biology falls squarely within the life sciences, restricting to citations with PubMed IDs is does not have a large effect on citation counts.

[6] https://help.altmetric.com/support/solutions/articles/6000190631-using-altmetric-data-for-altmetrics-research

publication (Campbell 1998). Another reason that release may come earlier than publication is because of the policy that all data is released after one year. If a team takes more than one year to publish results after the deposit, they would be forced to release at the one year point even if they eventually publish. Release sometimes happens after publication, but these cases should be rare and only be delayed for a few weeks. Any longer delays for release is either due to data errors or non-compliance with PDB policies.

Overall, 49 percent of the release dates are within two weeks of publication. This may lead to concerns about potential measurement error in the definition of the priority ordering. Throughout the paper, we always define the order of PDB release as the rule for being scooped. The community tracks public PDB releases carefully, so we believe this is a valid definition of priority. Publication dates are also complicated in recent years by the practice of online publication, which sometimes comes weeks before the print edition is published. But even if we prefer to consider only the publications as a claim to priority, our release date definition appears to usually correspond to the publication date ordering. In the 99 races where we have journal publication dates for the winner and loser, the priority ordering as defined by deposit corresponds with the priority ordering as defined by publication 83 percent of the time. To the degree that this is interpreted as measurement error, the scooped estimate will be somewhat attenuated.

A.5 Affiliations and University Rankings

Affiliation data is available from PubMed for most PDB deposits that resulted in a publication. Often the affiliation is only available for the first author of those publications, so we assign that affiliation to all authors on the publication. This assumption is more reasonable in structural biology than it is in economics for example, because cross-university collaboration is somewhat unusual in lab-based life sciences. The affiliations are contained in an author- or journal-reported text field that sometimes contains addresses or non-standard abbreviations. We standardize as many of these affiliations as possible using regular expressions and hand classification. We also assign as many affiliations as possible to their continent (Asia, North America, Europe, and other) to use as control variables. Affiliations are also categorized based on whether the affiliation is a university, non-profit research entity, or private corporation (typically a pharmaceutical company). In our full sample of projects (both racing and non-racing), there are 44,141 unique PubMed articles linked to the deposits. Of those papers, we were able to classify 71 percent to a standardized affiliation.

We link the university affiliations to the QS Top Universities Ranking for Life Sciences and Medicines.[7] This website provides rankings for 500 top academic programs based on surveys of academics and employers as well as citations per paper and h-index of the scientists affiliated with each department.

A.6 Name Disambiguation and Linked Author Papers in the PDB and PubMed

At various points in our analysis, we construct panel data of individual scientist and team productivity. First, we use measures of past PDB and PubMed productivity as control variables (Tables 3 and 4) and to predict citations as a measure of team reputation (Figures 8 and 9). Second, we use a panel of publications to construct long-run outcomes in the years following a scoop event (Table 5). The PDB does not explicitly link authors between deposits, and neither PubMed nor our version of Web of Science have author identifiers across publications. A further challenge is that many PDB deposits are not linked to a publication, so

[7] https://www.topuniversities.com/university-rankings/university-subject-rankings/2018/life-sciences-medicine

constructing control variables of past productivity is difficult using only publication data. We therefore use two separate approaches for constructing author-level panel variables: 1) Link PDB deposits by simple author name matching for control variables, 2) Use name disambiguation clustering from the Author-ity project (Torvik et al. 2005; Torvik and Smalheiser 2009) to count future publications and citations for long-run outcomes.

Simple Author Name Matching in PDB

In the first approach, we manually create a panel of author deposits and PDB-linked publications by matching last names and initials within the PDB. This name disambiguation procedure requires making assumptions about match reliability, and we follow the suggestions of Milojević (2013). We don't use additional information such as affiliations because they often change throughout a career, and are often only available for one author in the team.

The name disambiguation procedure using only last names and initials is more reliable in a smaller subset of academic papers. We therefore choose to focus the panel only on PubMed papers that are linked to the PDB instead of trying to use the full PubMed archive, which covers all of the medical and life science literature. This choice improves the reliability of our name-matching, but offers less information about academic productivity. Since we can use PDB name matching for unpublished deposits, we use this approach for constructing control variables for our main analysis.

Scientists usually identify themselves on publications with a consistent last name, but are sometimes inconsistent with their use of first and last initials, or first names and nicknames.[8] According to Milojević (2013), there are two potential matching errors that should be accounted for. First, a given individual may be identified as two or more authors (splitting). Second, two or more individuals may be identified as a single author (merging). We follow the hybrid model they propose to deal with these concerns, using first and second initials to determine whether splitting or merging is likely, especially in cases of very common last names.

To connect names across PDB-linked publications, we use the following procedure:

1. Strip names of non-alphabetic characters and standardize spacing and hyphenation of compound last names.
2. Identify groups of paper-authors that have the same last name and first initial.
3. Look at the second initial to determine potential merging errors. We find that 96.5 percent of the last name/first initial groups have no second-initial conflict, so we treat these as distinct individuals
4. If we are unable to differentiate the individual using the second initial, (e.g. JACKSON, P; JACKSON, PA; and JACKSON, PS), we keep them as a merged name, but mark the group as "common." These make up 3.5 percent of the sample.
5. We include a dummy control variable throughout the analysis that indicates the common names to help account for the possibility that name-matching errors are correlated with treatment.

We also use this panel to assign university rank and location controls. Racing projects sometimes go unpublished, so we cannot use the PDB-linked publication affiliation as a control variable in the main regression.

[8] Changes from maiden names to married names is also a potential source of error which we cannot account for, but this is becoming less common in recent years, especially among academics.

Therefore we assign the most recent affiliation of the first author in the publication panel to improve the coverage of these control variables.

Author-ity Name Disambiguation

For long-run productivity outcomes, we focus on a broader set of PubMed publications. For most authors, structural biology in the PDB is only one part of their scientific portfolio. Since simple name matching is not reliable in the full sample of PubMed publications, we use a dataset called Author-ity (Torvik et al. 2005; Torvik and Smalheiser 2009) to help disambiguate names. The Author-ity project is a large-scale, data-driven effort that incorporates additional information about co-author networks and research topics to separate unique authors within the full PubMed database. Each iteration of an author last name and first initial that appears on a PubMed paper is grouped together with the other papers that the algorithm infers to be the same individual and is assigned a unique person ID. For example, the name JACKSON, P has 293 different person IDs in Author-ity, each with a distinct set of PubMed identified papers.

If all PDB deposits were published, we could simply link the PDB deposits to the associated authors using PubMed IDs. But many of the racing projects are not published, so we need to match PDB author names to Author-ity name clusters and determine which cluster the PDB author belongs to. We first merge the full list of PDB author names to Author-ity using last name and first initial. We then mark every instance where a PDB-linked PubMed ID matches to a PubMed ID cluster within the Author-ity merged name.

These two steps leave us with three distinct groups of author names in the PDB:

1. Names that do not match to any Author-ity cluster (12 percent of racing sample authors). These are individuals who deposit at least once in the PDB, but never publish a paper (e.g. a graduate student that does not pursue academia).

2. Names that have PubMed IDs that match to one and only Author-ity person ID (60 percent of racing sample authors). We take this exclusive matching as evidence that all instances of the name in the PDB is a single person that is represented by the matched Author-ity person ID.

3. Names that have PubMed IDs that match to multiple Author-ity person IDs (29 percent of racing sample authors). These are common names that are likely distinct people within the PDB. We drop them from the long run analysis sample because we cannot determine which person is the author of a structure deposit that is not published.

We restrict our long-run analysis sample to the first two groups listed above (71 percent of racing sample authors). In this sub-sample, the individuals either never published a PubMed paper, or if they did, we have confidence that the PDB name represents a single individual.

Although our name disambiguation methods are not perfect, we rely on the assumption that any biases in our measures are equally distributed across winning and losing teams in a race. Given the balance in team characteristics shown in Table 3, we believe the winning teams are no more likely to have common names or mis-calculated productivity variables than losing teams, which should limit potential bias. To the extent that any remaining name matching mistakes create classical measurement error in the right-hand side variables, it would attenuate our results.

Supplemental Material (not copyedited or formatted) for: Ryan Hill, Carolyn Stein. 2025. "Scooped! Estimating Rewards for Priority in Science." Journal of Political Economy 133(3). DOI: https://doi.org/10.1086/733398.

B Protein Similarity and Race Definition

In this section we describe in detail the algorithm used to construct priority races used for our main analysis. Although the main text of the paper describes the basic rules for this sample construction, we report here a number of technical details and decisions that were used to construct the races in practice.

B.1 Sequence Similarity Algorithm

Each protein in the PDB is a chain composed of the 22 different types of proteinogenic amino acids in some combination. The order of these molecules in the chain defines the type of protein, and we use this code to compare the similarity of the proteins that scientists are working on. The PDB provides a clustering algorithm called the Basic Local Alignment Search Tool or BLAST (Altschul et al. 1990) which creates groupings of structure deposits that have identical or similar amino acid chains. The clusters can be defined at different thresholds of similarity, including 100 percent, 90 percent, 70 percent, and 50 percent. One possible approach to defining races would be to only focus on competing projects that determine the structure of proteins that are 100 percent similar. But in many cases, two proteins that are 90 percent similar or lower have many of the same defining features and functions within the same organism or across different species. Therefore, many interesting priority races are between teams working on very similar if not identical proteins. Following the similarity threshold chosen by (Brown and Ramaswamy 2007), we define racing for proteins all the way down to 50 percent similarity. We include races with a broad threshold in part to increase the sample size for our regressions, but also to include races over discoveries that were exceedingly different from any past structure discoveries. Other recent economics papers that study protein clusters also use similar cutoffs (Kim, 2023; Zhuo, 2022). We further validate our choice of similarity by comparing pairs of papers in each similarity category. Figure A1 calculates the share of scooped papers that cite the winning paper and plots it separately by sequence similarity. These are constructed as mutually exclusive groups with structures placed in the highest similarity cluster they appear in together. 70 percent, 90 percent, and 100 percent show almost identical rates, and the 50 percent similar pairs have only a slightly lower rate. Figure A2 shows the similarity between the winning and losing paper titles calculated with a character-replace string similarity metric. Here we see that the titles are equally similar between 50 percent, 70 percent, 90 percent, and 100 percent similarity papers.

Another tricky feature of the PDB data is that cluster groupings are sometimes defined at a level of granularity that is smaller than our outcome variables, which are defined at the structure deposit and article level. Proteins are composed of "chains" of amino acids, and large proteins are often characterized in the PDB as a set of distinct chains. Further, chains of amino acids are often grouped as "entities", and many proteins are combinations of two or more entities. This is relevant to our sample construction because the BLAST similarity algorithm clusters at the entity level rather than the protein level. In simple cases where proteins are made of a single entity (79 percent of structures in the PDB), a new structure discovery might directly scoop another team working on the same entity. But in some cases, a team working on a single entity might scoop a team that is working on a complex protein with multiple entities, only one of which was being worked on by both teams. These deposits will still be linked by the algorithm, but the interpretation of the scooping event is less obvious. We consider these cases to be "partial scoops" where some part of the scientific discovery was overshadowed by the winning team. Since outcomes are defined at the protein and paper level, including these partial scoops will potentially understate the effect of an average "full scoop." In our final regression sample, 68 percent of races are composed only of single-entity structures, 16 percent are exclusively multi-entity structures, and 16 percent are a mix of single- and multi-entity structures. We drop some very large proteins (such as the ribosome) that have more than 15 entities (0.7 percent of the sample). In these cases, the notion of a partial scoop is hard to define, as many different discoveries overlap at the entity level in sometimes complicated directions.

B.2 Procedure for defining races and scoop events

We follow the steps below to define priority races and scoop events. These steps are performed separately for four different similarity thresholds (50 percent, 70 percent, 90 percent, and 100 percent) and then combined in a final step.

1. Keep all clusters that have at least two deposits.

2. Sort the deposits within the clusters by release date, starting with the project that was released earliest. We focus only on cases of novel structure discoveries, so winners must be the first structure release in a given similarity cluster. We call this the priority deposit.

3. Compare the list of structure authors on the priority deposit with the list of authors on all subsequent deposits. Drop any follow-on deposits with one or more author names that were also on the priority deposit.[9]

4. Drop all deposits with a deposit date after the release date of the priority deposit. This rule allows for multiple teams to be scooped by the same priority structure. See Section 2.3 for a discussion of this rule.

This procedure identifies a set of races that are defined within 50 percent, 70 percent, 90 percent, or 100 percent similarity clusters. We consolidate to a final analysis sample that minimizes duplicate races and duplicate deposits. Using this procedure leaves us with some proteins that are scooped at multiple levels. For example, protein A may be first and protein B may be second in a 100 percent similar cluster but are also the first and second in a 90 percent similar cluster (and 70 percent and 50 percent). To avoid counting this race multiple times, we keep only the instance defined in the 100 percent sample. In more complicated cases, protein A might be scooped by protein B that is 70 percent similar, but also scooped by protein C that is 100 percent similar either before or after protein B is released. In these cases, we always keep the scoop event at the closest similarity. So the race between protein A and protein B is dropped, and the race between protein A and protein C is kept. This leaves us with a final sample of mutually exclusive races where each scooped paper only appears once. Some winning deposits are allowed to scoop more than one protein, sometimes at different similarity levels. In Appendix Table A5, we include robustness results of our main effects for races defined at the 100 percent level, and show that the results are comparable.

[9] In a few cases, we see instances where the same team of authors deposited multiple structure discoveries in the same cluster around the same time. We keep only one of those structures per team and give preference to the first deposit that resulted in a publication or the first one deposited if they are never published.

C Theoretical Appendix

C.1 A Model of Strategic Responses to Getting Scooped Before Project Completion

C.1.1 Setup

We start by considering the optimization problem of a scientist at the outset of a race with no information of her competitor's progress relative to her own. First, she chooses a maturation period m, which is the time she will spend on the project from start to finish. Higher m increases the value $V(m)$ of the project but also increases the cost $c \cdot m$ of the project, where $V(0) = 0$, $V'(m) > 0$, and $V''(m) < 0$. Given the utility function

$$u(m) = V(m) - c \cdot m$$

the scientist will select some m^* that maximizes her utility.[10] This m^* is defined by the first-order condition:

$$V'(m) = c.$$

As structural biology is a secretive field, we assume she has no knowledge about any potential competitors or their progress, and thus she will commit to this m^* unless additional information is revealed.

Now suppose there is a penalty for being scooped $\theta \in [0, 1]$ such that the losing team gets $\theta V(m)$ rather than $V(m)$ if the other team finishes first. Due to the attribution frictions raised by Dasgupta and David (1994) and the empirical evidence we show in Figure 6, we let the scoop penalty vary with the gap between when the two papers are released. In particular, let this gap be denoted $g(m)$ and let the scoop penalty be decreasing and convex in g. In other words, $\theta(0) = 1$, $\theta'(g) < 0$, and $\theta''(g) > 0$

C.1.2 Re-optimization.

If the scientist learns that she has been scooped before completing the project, new information is revealed and she has a chance to re-optimize. In other words, if she learns that she has been scooped at time $m_1 < m^*$, she now maximizes

$$u(\tilde{m}) = \theta(g(\tilde{m}))V(\tilde{m}) - \tilde{c} \cdot (\tilde{m} - m_1) \quad \text{subject to } \tilde{m} \geq m_1.$$

We let $\tilde{m}$ represent her new choice of m, after this revelation of information. $\theta(g(\tilde{m}))V(\tilde{m})$ is the value of the project despite getting scooped, and $\tilde{c} \cdot (\tilde{m} - m_1)$ is the remaining costs left to pay. In this case, $g(\tilde{m}) = \tilde{m} - m_1$. We assume that $\tilde{c} \leq c$, capturing the fact that the release of the first project may make some aspects of the project easier, due to informational spillovers. The solution to this optimization problem is $\tilde{m}^*$, which is implicitly defined by the first-order condition:

$$\theta(g(\tilde{m}^*))V'(\tilde{m}^*) = -\theta'(g(\tilde{m}^*))V(\tilde{m}^*) + \tilde{c}.$$

The left side of this equation represents the marginal benefit of increasing m: the marginal benefit of adding value. The right hand side represents the marginal costs of increasing m: the marginal decay in credit plus the marginal costs of spending additional time on the project. The behavior of both sides of the equation

[10] Note that to keep things simple, the researcher does not consider the probability of getting scooped in her selection of m. We could modify the problem to make $u(m)$ depend on some expectation of the credit she gets for $V(m)$ (similar to Hill and Stein (2024)). All we require here is a non-zero solution to the maximization problem.

depends on the specific functional forms of $\theta(\cdot)$ and $V(\cdot)$. However, we would like to show that it is possible for either $\tilde{m}^* < m^*$ or $\tilde{m}^* > m^*$. We can simply show that this is the true via specific examples.

Let $V(m) = \ln(1+m)$, and let $c = 0.5$. In this case, before knowledge of the scoop, the first-order condition yields:

$$V'(m^*) = c \implies \frac{1}{1+m^*} = 0.5 \implies m^* = 1.$$

Now, suppose that $\theta(g) = e^{-0.1g}$. To keep things simple, further assume that $m_1 = 0$, so that $\theta(g) = \theta(\tilde{m}) = e^{-0.1\tilde{m}}$. Start by letting $\tilde{c} = 0.5 = c$. The first-order condition for this problem yields:

$$\begin{aligned}\theta(\tilde{m}^*)V'(\tilde{m}^*) &= -\theta'(\tilde{m}^*)V(\tilde{m}^*) + \tilde{c}\\ e^{-0.1\tilde{m}^*}\left(\frac{1}{1+\tilde{m}^*}\right) &= 0.1e^{-0.1\tilde{m}^*}\ln(1+\tilde{m}^*) + 0.5\\ \tilde{m}^* &\approx 0.70 < m^*\end{aligned}$$

In this case, upon learning she has been scooped, the researcher will publish earlier than she originally planned. However, if instead $\tilde{c} = 0.25 < c$ due to informational spillovers from the first project, then solving the first-order condition yields $\tilde{m}^* \approx 1.58 > m^*$. In this case, upon learning she has been scooped, the researcher decides to slow down.

In general, we have three possible cases:

1. Case 1 (“hurry up and finish”): In this case, the researcher chooses $\tilde{m}^* < m^*$, because the decaying credit plus the continuation costs outweigh increasing the value of the project.

2. Case 2 (“delay and expand”): In this case, the researcher chooses $\tilde{m^*} > m^*$, because the increasing value of the project outweighs the decaying credit plus continuation costs.

3. Case 3 (“abandon”): Of course, it is possible that even after selecting a new $\tilde{m}^*$, the benefits once the researcher knows they will be scooped are not enough to outweigh the costs of completing the project. These projects will therefore go unfinished.

C.2 A Model of Academic Attention

C.2.1 Setup

Editors, reviewers, and authors read new academic papers. In doing so, they receive a noisy signal of the paper’s quality. The notion that paper quality is only partially observed by readers is similar to the setup in Card and DellaVigna (2020) and may arise from inattention or uncertainty about the importance of the contribution. The signal, s, is a function of the paper’s true underlying quality (q) as well as a noise term, u:

$$s = q + u$$

where $u \sim N(0, \sigma_u^2)$ is independent of $q \sim N(\alpha, \sigma_q^2)$. Following the standard statistical discrimination model, readers will use both the signal and the average quality to infer the paper’s quality:

$$\hat{q}(s) = E[q|s] = \lambda s + (1-\lambda)\alpha$$

where $\lambda = \frac{\sigma_q^2}{\sigma_q^2+\sigma_u^2}$ is the signal-to-noise ratio. Intuitively, expected quality is a weighted average of the observed signal and mean quality. Readers put more weight on the signal when λ is large, i.e. when the signal is informative relative to the noise term.

The Priority Premium When making decisions about which paper to publish or cite, scientists care about both quality and priority. Consider two papers which answer the same question, with inferred qualities $\hat{q}_1$ and $\hat{q}_2$. Let the numeric subscript index the order of publication, so that $\hat{q}_1$ was published before $\hat{q}_2$, and let $f > 0$ denote the priority premium. A scientist will cite the first paper if $\hat{q}_1 + f \geq \hat{q}_2$. On the other hand, a scientist will cite the second paper if $\hat{q}_1 + f < \hat{q}_2$.

Lab Types Suppose there are two types of labs, H and L. H labs are "high-reputation" labs, known for producing papers of high average quality, while L labs are "low-reputation" labs, known for producing papers of low average quality. In other words, q is drawn from a different distribution depending on the lab type. For H labs, $q^H \sim N\left(\alpha^H, \sigma_q^2\right)$ while for L labs, $q^L \sim N\left(\alpha^L, \sigma_q^2\right)$. The key distinction between the two lab types is that $\alpha^H > \alpha^L$. We will assume that variances are equal.

When two labs each write a paper on the identical topic (or in our case, protein), the true qualities of the two papers are the same. However, if the labs have different reputations, the inferred qualities will be different, even if the signals are identical:

$$\hat{q}^H(s) = \lambda s + (1-\lambda)\alpha^H$$
$$\hat{q}^L(s) = \lambda s + (1-\lambda)\alpha^L.$$

Ultimately, this gives rise to two distinct effects when competing labs publish on the same protein. The "priority effect" leads scientists to cite the earlier paper, since this paper receives a premium, as described above. On the other hand, the "reputation effect" leads scientists to cite the paper from the higher-reputation lab, since this paper will have higher inferred quality. This insight leads us to two propositions.

Proposition 1. *If labs are the same type, then the lab that publishes first is more likely to be cited. In other words,*

$$P(\hat{q}_1^H + f \geq \hat{q}_2^H) = P(\hat{q}_1^L + f \geq \hat{q}_2^L) > \frac{1}{2}.$$

Proof of Proposition 1.

The intuition is that if the labs are the same type, there is no differential reputation effect. Therefore, citations are driven solely by the priority effect. Consider two high-reputation labs, H_1 and H_2. H_1 publishes before H_2. The probability that H_1 is cited is:

$$\begin{aligned}
P\left(\hat{q}_1^H + f > \hat{q}_2^H\right) &= P\left((1-\lambda)\alpha^H + \lambda s_1 + f > (1-\lambda)\alpha^H + \lambda s_1\right) \\
&= P\left(\lambda(q+u_1) + f > \lambda(q+u_2)\right) \\
&= P\left(\lambda u_1 + f > \lambda u_2\right) \\
&= P\left(u_2 - u_1 < \frac{f}{\lambda}\right) \\
&= P\left(\frac{u_2 - u_1}{\sqrt{2}\sigma_u} < \frac{f}{\lambda\sqrt{2}\sigma_u}\right) \\
&= \Phi\left(\frac{f}{\lambda\sqrt{2}\sigma_u}\right) \\
&> \frac{1}{2}
\end{aligned}$$

using the fact that $(u_2 - u_1) \sim N\left(0, 2\sigma_u^2\right)$ and $f, \lambda > 0$. Similarly, consider two low-reputation labs, L_1 and L_2. L_1 publishes before L_2. Analogously, the probability that L_1 is cited is $\Phi\left(\frac{f}{\lambda\sqrt{2}\sigma_u}\right) > \frac{1}{2}$.

Proposition 2. *If the lab that publishes first is H-type and the lab that publishes second is L-type, then the lab that publishes first is more likely to be cited. Moreover, the difference in citations will be greater than if the labs were the same type. Conversely, if the lab that publishes first is L-type and the lab that publishes second is H-type, it is ambiguous which lab is more likely to be cited. However, the difference in probability of citation will certainly be less than if the labs were the same type. This means that we can rank the probability of citation in all four scenarios:*

$$P(\hat{q}_1^H + f \geq \hat{q}_2^L) > P(\hat{q}_1^H + f \geq \hat{q}_2^H) = P(\hat{q}_1^L + f \geq \hat{q}_2^L) > P(\hat{q}_1^L + f \geq \hat{q}_2^H).$$

Proof of Proposition 2.

The intuition is that if the first lab is H-type and the second lab is L-type, then the priority effect and the reputation effect work in the same direction. However, if the first lab is L-type and the second lab is H-type, then the priority effect and the reputation effect are working in opposite directions. Therefore, the net effect on citation behavior is ambiguous.

Consider a high-reputation lab and a low-reputation lab, H_1 and L_2. H_1 publishes before L_2. The probability that H_1 is cited is:

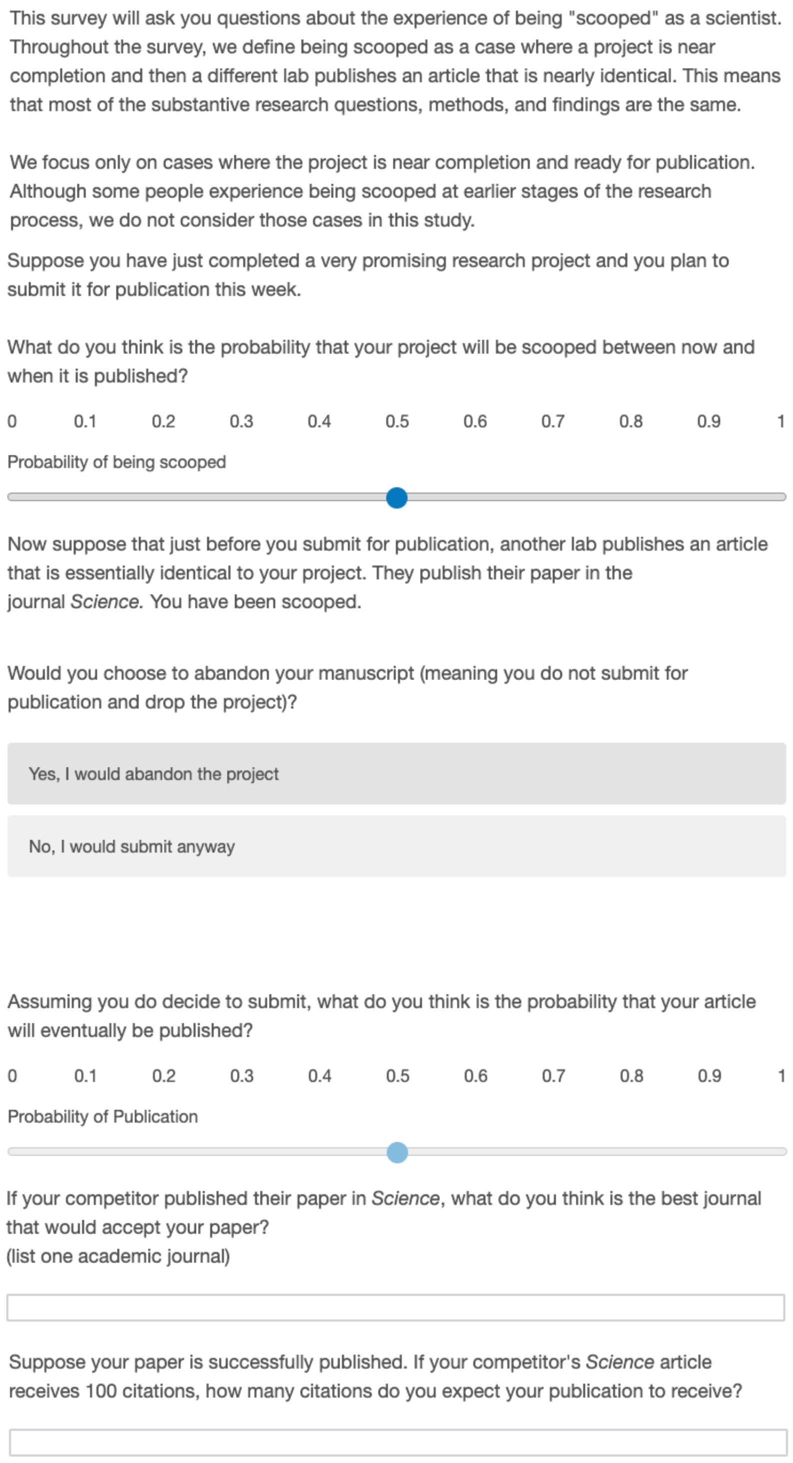
$$\begin{aligned}
P\left(\hat{q}_H + f > \hat{q}_L\right) &= P\left((1-\lambda)\alpha^H + \lambda s_1 + f > (1-\lambda)\alpha^L + \lambda s_2\right) \\
&= P\left((1-\lambda)\alpha^H + \lambda(q+u_1) + f > (1-\lambda)\alpha^L + \lambda(q+u_2)\right) \\
&= P\left((1-\lambda)(\alpha^H - \alpha^L) + f > \lambda(u_2 - u_1)\right) \\
&= P\left(u_2 - u_1 < \frac{(1-\lambda)(\alpha^H - \alpha^L) + f}{\lambda}\right). \\
&= P\left(\frac{u_2 - u_1}{\sqrt{2}\sigma_u} < \frac{(1-\lambda)(\alpha^H - \alpha^L) + f}{\lambda\sqrt{2}\sigma_u}\right) \\
&= \Phi\left(\frac{(1-\lambda)(\alpha^H - \alpha^L) + f}{\lambda\sqrt{2}\sigma_u}\right) \\
&> \Phi\left(\frac{f}{\lambda\sqrt{2}\sigma_u}\right) > \frac{1}{2}
\end{aligned}$$

again using the fact that $(u_2 - u_1) \sim N\left(0, 2\sigma_u^2\right)$ and $(1-\lambda) > 0$, $\alpha_H > \alpha_L$. Similarly, consider a low-reputation lab and a high-reputation lab, L_1 and H_2. L_1 publishes before H_2. The probability that L_1 is cited is:

$$\begin{aligned}
P\left(\hat{q}_L + f > \hat{q}_H\right) &= P\big((1-\lambda)\alpha^L + \lambda s_1 + f > (1-\lambda)\alpha^H + \lambda s_2\big) \\
&= P\big((1-\lambda)\alpha^L + \lambda(q+u_1) + f > (1-\lambda)\alpha^H + \lambda(q+u_2)\big) \\
&= P\big(-(1-\lambda)(\alpha^H - \alpha^L) + f > \lambda(u_2 - u_1)\big) \\
&= P\left(u_2 - u_1 < \frac{-(1-\lambda)(\alpha^H - \alpha^L) + f}{\lambda}\right). \\
&= P\left(\frac{u_2 - u_1}{\sqrt{2}\sigma_u} < \frac{-(1-\lambda)(\alpha^H - \alpha^L) + f}{\lambda\sqrt{2}\sigma_u}\right) \\
&= \Phi\left(\frac{-(1-\lambda)(\alpha^H - \alpha^L) + f}{\lambda\sqrt{2}\sigma_u}\right) \\
&< \Phi\left(\frac{f}{\lambda\sqrt{2}\sigma_u}\right).
\end{aligned}$$

Whether the expression is greater or less than $\frac{1}{2}$ depends on the magnitude of $(1-\lambda)(\alpha^H - \alpha^L)$. More specifically, if $(1-\lambda)(\alpha^H - \alpha^L) < f$, then $P\left(\hat{q}_L + f > \hat{q}_H\right) > \frac{1}{2}$. If $(1-\lambda)(\alpha^H - \alpha^L) > f$, then $P\left(\hat{q}_L + f > \hat{q}_H\right) < \frac{1}{2}$.

D Survey Text



This survey will ask you questions about the experience of being "scooped" as a scientist. Throughout the survey, we define being scooped as a case where a project is near completion and then a different lab publishes an article that is nearly identical. This means that most of the substantive research questions, methods, and findings are the same.

We focus only on cases where the project is near completion and ready for publication. Although some people experience being scooped at earlier stages of the research process, we do not consider those cases in this study.

Suppose you have just completed a very promising research project and you plan to submit it for publication this week.

What do you think is the probability that your project will be scooped between now and when it is published?

Now suppose that just before you submit for publication, another lab publishes an article that is essentially identical to your project. They publish their paper in the journal *Science.* You have been scooped.

Would you choose to abandon your manuscript (meaning you do not submit for publication and drop the project)?

Assuming you do decide to submit, what do you think is the probability that your article will eventually be published?

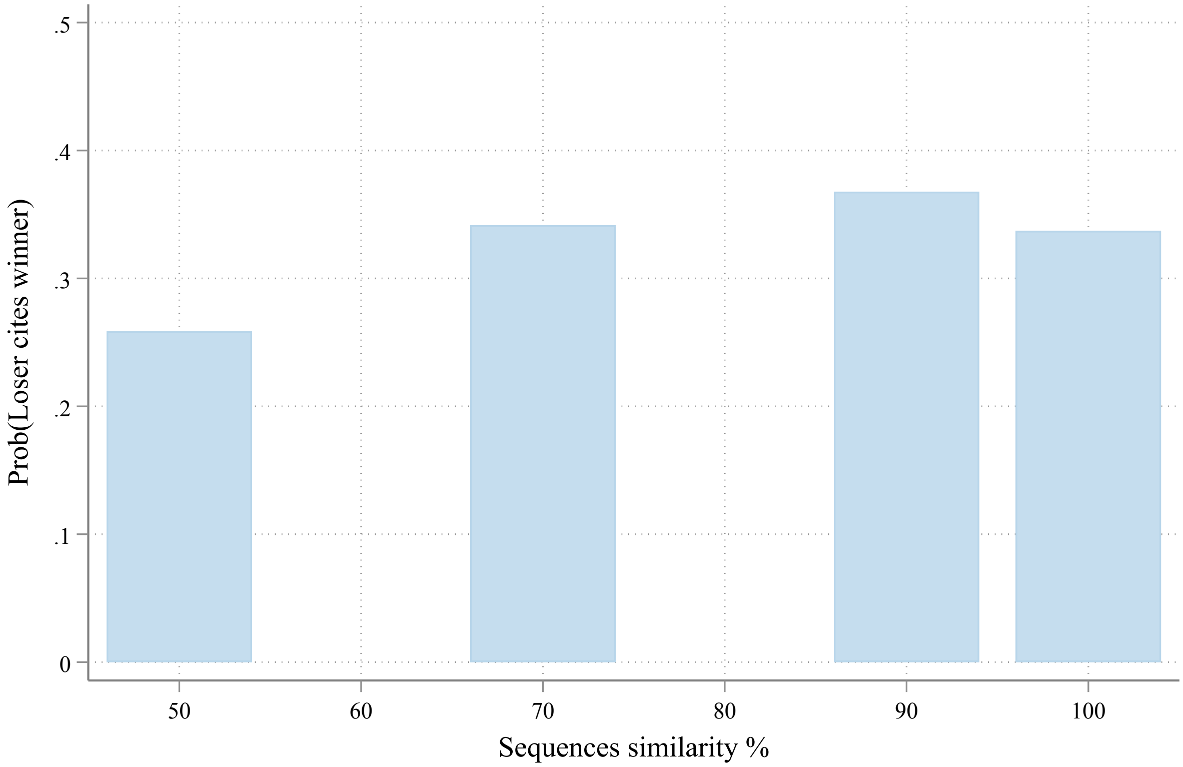
If your competitor published their paper in *Science*, what do you think is the best journal that would accept your paper?
(list one academic journal)

Suppose your paper is successfully published. If your competitor's *Science* article receives 100 citations, how many citations do you expect your publication to receive?

Supplemental Material (not copyedited or formatted) for: Ryan Hill, Carolyn Stein. 2025. "Scooped! Estimating Rewards for Priority in Science." Journal of Political Economy 133(3). DOI: https://doi.org/10.1086/733398.

E Appendix Figures and Tables

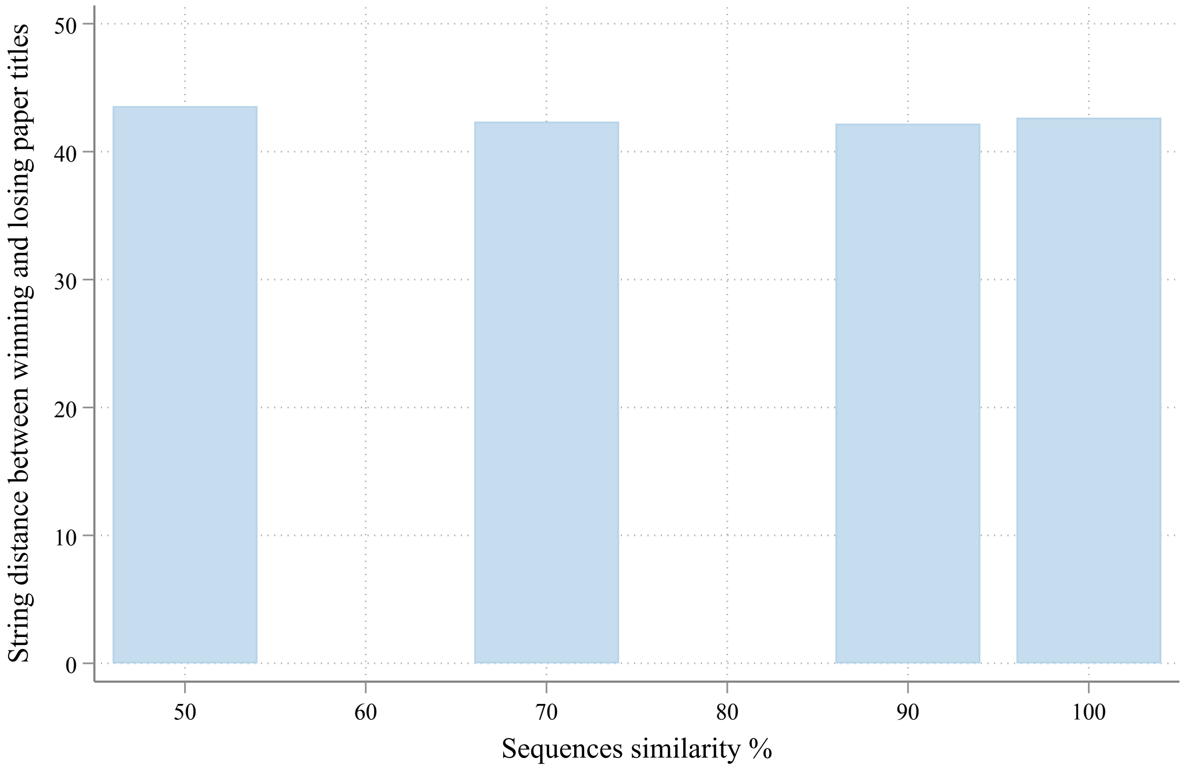
Figure A1: Probability of Loser Citing Winner by Sequence Similarity



Notes: This figure shows the probability that the losing team cites the winning team at increasing levels of sequence similarity. Similarity groups are mutually exclusive so that races are placed in the highest similarity cluster in which they appear together.

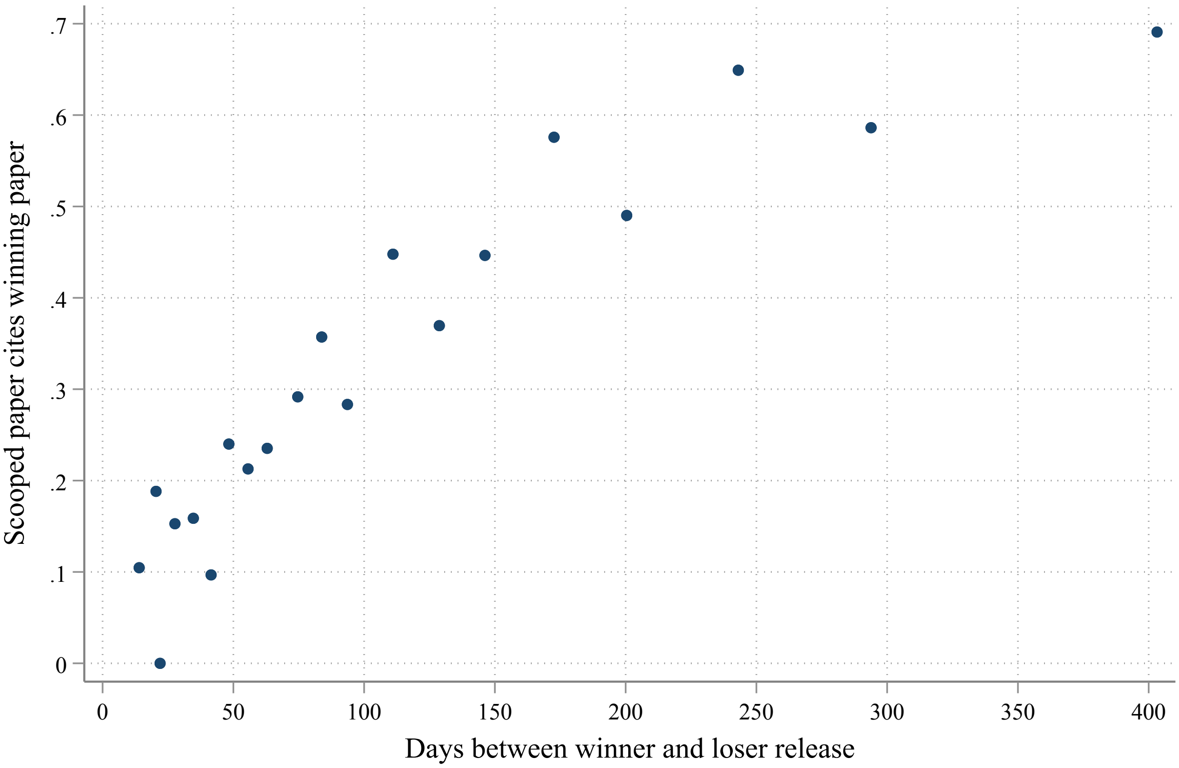
Supplemental Material (not copyedited or formatted) for: Ryan Hill, Carolyn Stein. 2025. "Scooped! Estimating Rewards for Priority in Science." Journal of Political Economy 133(3). DOI: https://doi.org/10.1086/733398.

Figure A2: Title Similarity Between Winning and Losing Paper by Sequence Similarity



Notes: This figure shows the character replacement string similarity of the titles of the winning and losing papers at increasing levels of sequence similarity. Similarity groups are mutually exclusive so that races are placed in the highest similarity cluster in which they appear together.

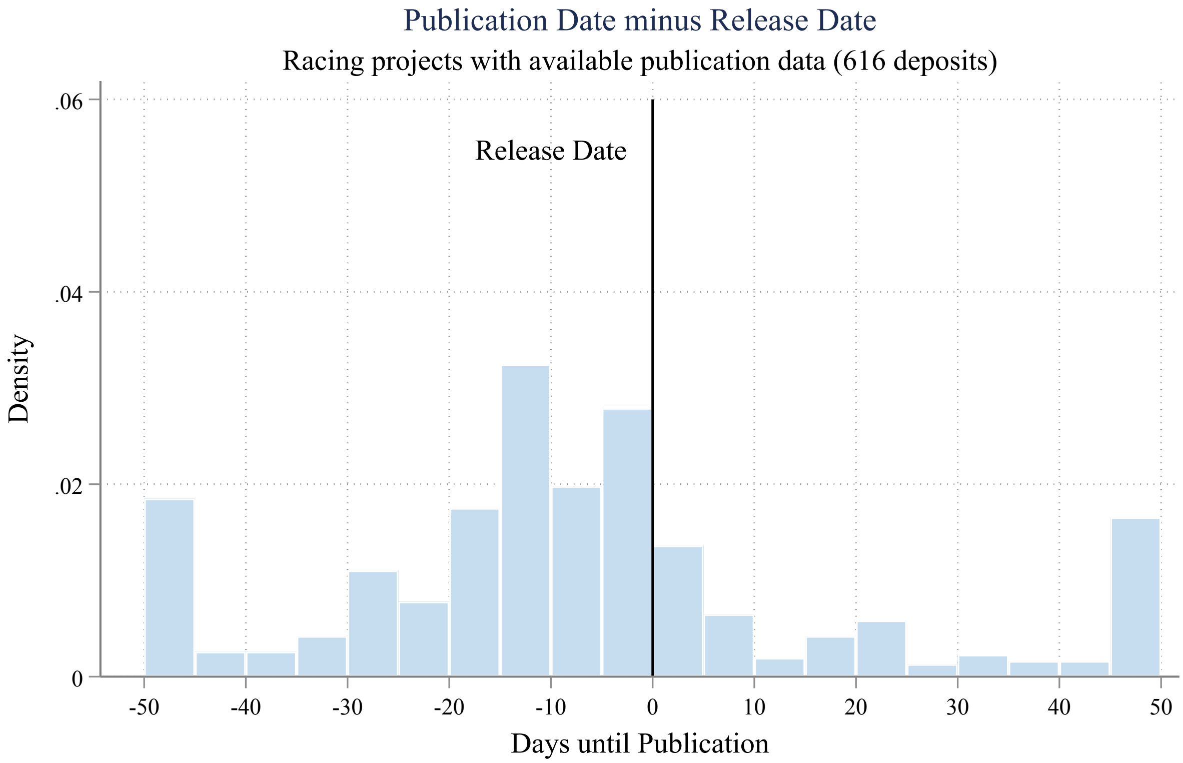
Figure A3: Probability that Scooped Paper Cites Winning Paper by Release Date Gap



Notes: This binned scatterplot shows the probability that the scooped paper cited the winning paper by the number of days between the release dates of the winning and losing projects. Sample is 1,149 races where both teams published papers with a PubMed ID.

Supplemental Material (not copyedited or formatted) for: Ryan Hill, Carolyn Stein. 2025. "Scooped! Estimating Rewards for Priority in Science." Journal of Political Economy 133(3). DOI: https://doi.org/10.1086/733398.

Figure A4: Correspondence Between Release Date and Available Publication Dates



Notes: This histogram shows the correspondence between PDB release date and publication date when publication dates are available from the editorial date supplement. Positive days means the publication came before release, and negative days mean it came after release.

Table A1: Lasso-selected Variables and Coefficients for Predicted Citations

Lasso-selected variables	Post-Lasso OLS coefficients
Number of authors	0.54
Affiliation in North America	1.72
Affiliation in Asia	-3.53
Non-academic affiliation	1.73
First author experience (years)	-0.20
First author top-5 publications, 5 prior years	2.45
First author PDB deposits, all years squared	0.00
First author PDB deposits, 5 prior years squared	0.00
First author publications, 5 prior years squared	0.00
Last author experience (years)	-0.22
Last author PDB deposits, 5 prior years	-0.11
Last author publications, 5 prior years	0.02
Last author top-5 publications, all years	0.21
Last author top-5 publications, 5 prior years	2.14
Last author PDB deposits, all years squared	0.00
Last author PDB deposits, 5 prior years squared	0.00
Last author top-10 publications, 5 prior years squared	-0.01
University rank bins:	
1-10	3.48
71-80	-0.17
81-90	-1.03
101-110	-2.39
111-120	5.03
151-160	-2.67
171-180	-2.08
181-190	-0.40
211-220	-5.18
231-240	-1.43
271-280	-4.11
291-300	-2.85
401-410	-2.70
Constant	10.28
R-squared	0.102
N	58,758

Notes: This table presents results from a Lasso regression of 3-year unconditional citations on observable team characteristics. The model is estimated in the non-racing sample and uses data-driven and heteroskedasticity-robust penalization. Estimated coefficients are from a post-Lasso OLS regression of 3-year citations on selected regressors.

Supplemental Material (not copyedited or formatted) for: Ryan Hill, Carolyn Stein. 2025. "Scooped! Estimating Rewards for Priority in Science." Journal of Political Economy 133(3). DOI: https://doi.org/10.1086/733398.

Table A2: Effect of Getting Scooped on Project Outcomes - Oster (2019) Robustness Check

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	Five-year citations (4)	Top-10% five year citations (5)
Panel A. No controls, no FE					
Scooped	-0.025**	-0.191***	-0.064***	-0.239***	-0.035***
	(0.011)	(0.031)	(0.014)	(0.051)	(0.010)
	[0.001]	[0.008]	[0.005]	[0.006]	[0.003]
Panel B. Base controls, protein FE					
Scooped	-0.026**	-0.182***	-0.063***	-0.216***	-0.028**
	(0.013)	(0.045)	(0.021)	(0.063)	(0.014)
	[0.704]	[0.676]	[0.607]	[0.767]	[0.725]
Oster (2019) Bias-adjusted β	-0.027	-0.177	-0.061	-0.209	-0.025
Selection ratio (δ) needed for $\beta = 0$	60.0	15.2	14.0	16.5	7.9

Notes: This table presents regression estimates of the scoop penalty following equation 2 in the text (see Table 4). Panel A reports coefficients from a simple bivariate regression with no controls or protein fixed effects with standard errors in parentheses and R^2 in brackets. Panel B includes all base controls and protein fixed effects, comparable to panel B in Table 4. The Oster (2019) bias adjusted coefficient assumes a maximum $R^2 = 1$ and $\delta = 1$, meaning we assume that treatment is selected equally on observables and unobservables. The selection ratio (δ) needed for $\beta = 0$ shows that treatment would need to be 7 times more selected on unobservables than observables for the coefficient to equal zero.

**p<0.1, **p<0.05, ***p<0.01.*

Table A3: Effect of Getting Scooped on Project Outcomes - Alternative Hit Rate Metrics

Dependent variable	Top-1% three year citations (1)	Top-5% three year citations (2)	Top-10% three year citations (3)	Top-1% five year citations (4)	Top-5% five year citations (5)	Top-10% five year citations (6)	Top-1% ten year citations (7)	Top-5% ten year citations (8)	Top-10% ten year citations (9)
Panel A. No controls									
Scooped	-0.007	-0.023**	-0.033**	-0.006	-0.017*	-0.037***	-0.011*	-0.031**	-0.049***
	(0.005)	(0.009)	(0.013)	(0.005)	(0.010)	(0.014)	(0.006)	(0.014)	(0.017)
Panel B. Base controls									
Scooped	-0.007*	-0.020**	-0.027**	-0.005	-0.013	-0.028**	-0.009	-0.027*	-0.044***
	(0.004)	(0.010)	(0.013)	(0.005)	(0.010)	(0.014)	(0.006)	(0.014)	(0.017)
Panel C. PDS-Lasso selected controls									
Scooped	-0.007**	-0.022***	-0.031***	-0.005	-0.015**	-0.036***	-0.010**	-0.030***	-0.046***
	(0.003)	(0.007)	(0.010)	(0.003)	(0.007)	(0.010)	(0.004)	(0.010)	(0.012)
Winner Y mean	0.012	0.076	0.148	0.011	0.068	0.149	0.012	0.081	0.153
Observations	2,931	2,931	2,931	2,514	2,514	2,514	1,515	1,515	1,515

Notes: This table presents regression estimates of the scoop penalty with alternative hit rate measures, following equation 2 in the text. Each regression contains protein (i.e., race) fixed effects. Observations are at the structure level. Each coefficient is from a separate regression. Panel A presents results from a specification with no controls. Panel B adds the base set of controls as listed in Table 3. Panel C uses controls selected by the PDS-Lasso method. Standard errors are in parentheses, and are clustered at the race level.

**p<0.1, **p<0.05, ***p<0.01.*

Supplemental Material (not copyedited or formatted) for: Ryan Hill, Carolyn Stein. 2025. "Scooped! Estimating Rewards for Priority in Science."
Journal of Political Economy 133(3). DOI: https://doi.org/10.1086/733398.

Table A4: Effect of Getting Scooped on Project Outcomes - No Protein (i.e., Race) Fixed Effects

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	Five-year citations (4)	Top-10% five year citations (5)
Panel A. No controls					
Scooped	-0.025**	-0.191***	-0.064***	-0.239***	-0.035***
	(0.011)	(0.031)	(0.014)	(0.051)	(0.010)
Panel B. Base controls					
Scooped	-0.022**	-0.154***	-0.051***	-0.180***	-0.025**
	(0.009)	(0.032)	(0.014)	(0.046)	(0.010)
Panel C. PDS-Lasso selected controls					
Scooped	-0.021**	-0.146***	-0.058***	-0.169***	-0.024**
	(0.010)	(0.031)	(0.015)	(0.046)	(0.010)
Winner Y mean	0.879	-0.027	0.320	28.830	0.149
Observations	3,279	3,279	3,279	2,514	2,514

Notes: This table presents regression estimates of the scoop penalty, following equation 2 in the text, but excluding protein (i.e., race) fixed effects. Observations are at the structure level. Each coefficient is from a separate regression. Panel A presents results from a specification with no controls. Panel B adds the base set of controls as listed in Table 3. Panel C uses controls selected by the PDS-Lasso method. Standard errors are in parentheses, and are clustered at the race level. Column 4 regression uses asinh(five-year citations) as the dependent variable, but Winner Y Mean is reported in levels for ease of interpretation.
**p<0.1, **p<0.05, ***p<0.01.*

Table A5: Effect of Getting Scooped on Project Outcomes - 100 Percent Sequence Similarity

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	Five-year citations (4)	Top-10% five year citations (5)
Panel A. No controls					
Scooped	-0.023	-0.174**	-0.051	-0.272**	-0.047**
	(0.025)	(0.070)	(0.032)	(0.112)	(0.021)
Panel B. Base controls					
Scooped	-0.035	-0.156**	-0.044	-0.288***	-0.032
	(0.022)	(0.074)	(0.034)	(0.109)	(0.020)
Panel C. PDS-Lasso selected controls					
Scooped	-0.027	-0.172***	-0.049**	-0.253***	-0.047***
	(0.018)	(0.052)	(0.023)	(0.080)	(0.015)
Winner Y mean	0.882	-0.078	0.289	27.956	0.138
Observations	1,178	1,178	1,178	891	891

Notes: This table presents regression estimates of the scoop penalty comparable to Table 4 in the main text. This version restricts to protein clusters in which the BLAST algorithm classifies the protein sequences as being 100% similar. This sub-sample therefore offers the narrowest definition of a scoop where the racing projects are scientifically identical. See Table 4 notes for regression details.
**p<0.1, **p<0.05, ***p<0.01.*

Supplemental Material (not copyedited or formatted) for: Ryan Hill, Carolyn Stein. 2025. "Scooped! Estimating Rewards for Priority in Science." Journal of Political Economy 133(3). DOI: https://doi.org/10.1086/733398.

Table A6: Effect of Getting Scooped on Alternative Measures of Attention

Dependent variable: All transformed with asinh()	Mendeley downloads (1)	News stories (2)	Wikipedia citations (3)	Patent citations (4)	Twitter mentions (5)	Atltmetric attention (6)
Panel A. No controls						
Scooped	-0.468***	-0.108**	-0.038**	-0.009	-0.112	-0.246***
	(0.151)	(0.042)	(0.018)	(0.028)	(0.078)	(0.095)
Panel B. Base controls						
Scooped	-0.462***	-0.092**	-0.031	-0.003	-0.094	-0.216**
	(0.146)	(0.043)	(0.020)	(0.031)	(0.075)	(0.091)
Panel C. PDS-Lasso selected controls						
Scooped	-0.427***	-0.103***	-0.036***	-0.010	-0.095*	-0.228***
	(0.103)	(0.031)	(0.014)	(0.021)	(0.054)	(0.066)
Winner Y mean	43.025	0.650	0.105	0.262	3.974	9.201
Observations	1,321	1,321	1,321	1,321	1,321	1,321

Notes: Attention outcomes are sourced from Altmetric.com. Sample restricted to years 2011-2017. Each regression contains protein (i.e. race) fixed effects. Observations are at the structure level. Each coefficient is from a separate regression. Panel A presents results from a specification with no controls. Panel B adds the base set of controls as listed in Table 3. Panel C uses controls selected by the PDS-Lasso method. Standard errors are in parentheses, and are clustered at the race level. All outcomes are cumulative counts of the metrics summed over time between the publication date to August 2019. All counts are transformed with the inverse hyperbolic sine transformation. The Altmetric Attention Score is a composite measure of all metrics used by Altmetric.com.

**p<0.1, **p<0.05, ***p<0.01.*

Table A7: Effect of Getting Scooped on Three-Year Productivity

			Total count three years after race				
	Any PubMed	Any PDB	PubMed	PDB	Top-ten	Citation-weighted	Top-10% cited
	3 years later	3 years later	Publications	Publications	publications	publications	publications
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. All scientists							
Scooped	-0.012**	-0.039***	-0.340	-0.097	-0.037	-0.188***	-0.334**
	(0.006)	(0.011)	(0.520)	(0.117)	(0.061)	(0.042)	(0.133)
Winner Y mean	0.824	0.646	27.145	4.305	2.184	297.661	4.655
Observations	10,033	10,033	10,033	10,033	10,033	7,660	7,660
Panel B. Novices							
Scooped	-0.034**	-0.019	-0.044	-0.091	0.074*	-0.271***	-0.069
	(0.016)	(0.018)	(0.142)	(0.097)	(0.040)	(0.085)	(0.065)
Winner Y mean	0.428	0.309	2.307	1.097	0.334	44.063	0.680
Observations	2,369	2,369	2,369	2,369	2,369	1,806	1,806
Panel C. Veterans							
Scooped	-0.006	-0.037***	-0.184	-0.060	-0.060	-0.171***	-0.472**
	(0.004)	(0.013)	(0.794)	(0.163)	(0.089)	(0.047)	(0.191)
Winner Y mean	0.983	0.781	36.687	5.595	2.917	400.753	6.263
Observations	6,729	6,729	6,729	6,729	6,729	5,167	5,167

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 3 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 6 dependent variable is the total citations accrued in three years to all papers published in the three years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 7 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with seven years or less of publishing experience prior to the priority race year), and panel C restricts to veterans (defined as all non-novices). All regressions include scientist-level covariates selected by PDS-Lasso and race fixed effects. Standard errors are in parentheses, and are clustered at the race level.

**p<0.1, **p<0.05, ***p<0.01.*

Table A8: Effect of Getting Scooped on Ten-Year Productivity

			Total count ten years after race				
	Any PubMed	Any PDB	PubMed	PDB	Top-ten	Citation-weighted	Top-10% cited
	10 years later	10 years later	Publications	Publications	publications	publications	publications
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. All scientists							
Scooped	-0.012*	-0.034***	-2.786	0.042	-0.269	-0.026	-1.008
	(0.007)	(0.013)	(2.786)	(0.526)	(0.232)	(0.071)	(0.635)
Winner Y mean	0.857	0.739	91.647	13.965	7.090	928.013	14.076
Observations	5,351	5,351	5,351	5,351	5,351	3,114	3,114
Panel B. Novices							
Scooped	-0.044*	-0.056**	0.192	0.276	0.212	-0.168	0.410
	(0.022)	(0.026)	(0.803)	(0.470)	(0.183)	(0.148)	(0.338)
Winner Y mean	0.513	0.417	9.900	3.739	1.301	122.905	1.792
Observations	1,258	1,258	1,258	1,258	1,258	743	743
Panel C. Veterans							
Scooped	-0.002	-0.026*	-5.335	-0.873	-0.749**	-0.121*	-1.988**
	(0.002)	(0.013)	(4.058)	(0.710)	(0.333)	(0.065)	(0.875)
Winner Y mean	0.995	0.869	124.346	18.108	9.418	1243.954	18.891
Observations	3,607	3,607	3,607	3,607	3,607	2,079	2,079

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 3 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 6 dependent variable is the total citations accrued in three years to all papers published in the ten years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 7 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for all scientists. Panel B restricts to novices (defined as scientists with seven years or less of publishing experience prior to the priority race year), and panel C restricts to veterans (defined as all non-novices). All regressions include scientist-level covariates selected by PDS-Lasso and race fixed effects. Standard errors are in parentheses, and are clustered at the race level.

**p<0.1, **p<0.05, ***p<0.01.*

Table A9: Effect of Getting Scooped on Five-Year Productivity, Author Position Subsamples

			Total count within five years after race				
	Any PubMed within five years	Any PDB within five years	PubMed publications	PDB publications	Top-ten publications	Citation-weighted publications	Top-10% cited publications
Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. First Authors							
Scooped	-0.031*	-0.037	1.982	-0.025	0.018	-0.133	0.691
	(0.017)	(0.023)	(2.169)	(0.283)	(0.135)	(0.108)	(0.657)
Winner Y mean	0.821	0.692	31.576	4.191	2.045	278.251	4.296
Observations	1,166	1,166	1,166	1,166	1,166	890	890
Panel B. Middle Authors							
Scooped	-0.020**	-0.047***	-1.828	-0.279	0.013	-0.237***	-0.613**
	(0.010)	(0.016)	(1.428)	(0.237)	(0.129)	(0.071)	(0.298)
Winner Y mean	0.828	0.658	42.433	5.476	3.020	481.690	7.378
Observations	4,833	4,833	4,833	4,833	4,833	3,624	3,624
Panel C. Last Authors							
Scooped	-0.012	-0.044**	-2.515	-0.721	-0.775**	-0.311***	-1.099**
	(0.009)	(0.019)	(2.330)	(0.641)	(0.310)	(0.093)	(0.481)
Winner Y mean	0.901	0.843	61.543	14.557	6.629	669.102	10.964
Observations	1,190	1,190	1,190	1,190	1,190	900	900

Notes: This table presents regression estimates of the long-run scoop penalty, following equation 3 in the text. Observations are at the scientist level. Each coefficient is from a separate regression. Column 6 dependent variable is the total citations accrued in three years to all papers published in the five years after the race transformed with the the inverse hyperbolic sine function (winner Y means reported in level citations). Column 7 dependent variable is the total number of publications that reach the top-10% of three-year citations in that publishing year. Panel A presents results for the first scientist listed on the structure deposit, Panel B restricts to middle authors, and Panel C restricts to last authors. We use the the author list and ordering on the structure deposit because it is available for all teams regardless of publication status. It is usually the same as the resulting paper author list and ordering but with occasional differences. All regressions include scientist-level covariates selected by PDS-Lasso and race fixed effects. Standard errors are in parentheses, and are clustered at the race level.

**p<0.1, **p<0.05, ***p<0.01.*

Table A10: Effect of Getting Scooped Prior to Deposit

Dependent variable	Published (1)	Std. journal impact factor (2)	Top-ten journal (3)	Five-year citations (4)	Top-10% five year citations (5)
Panel A. No controls					
Scooped	0.023*	-0.156***	-0.072***	-0.136**	-0.038***
	(0.014)	(0.038)	(0.016)	(0.067)	(0.014)
Panel B. Base controls					
Scooped	-0.025**	-0.225***	-0.093***	-0.310***	-0.046***
	(0.011)	(0.040)	(0.016)	(0.061)	(0.015)
Panel C. PDS-Lasso selected controls					
Scooped	-0.020**	-0.216***	-0.087***	-0.284***	-0.042***
	(0.008)	(0.029)	(0.012)	(0.044)	(0.010)
Winner Y mean	0.842	-0.116	0.278	29.167	0.152
Observations	4,830	4,830	4,830	3,238	3,238

Notes: This table presents regression estimates of the scoop penalty restricting to scoops that occured prior to deposit. Each regression contains protein (i.e., race) fixed effects. Observations are at the structure level. Each coefficient is from a separate regression. Panel A presents results from a specification with no controls. Panel B adds the base set of controls as listed in Table 3. Panel C uses controls selected by the PDS-Lasso method. Standard errors are in parentheses, and are clustered at the race level. Column 4 regression uses asinh(five-year citations) as the dependent variable, but Winner Y Mean is reported in levels for ease of interpretation.
**p<0.1, **p<0.05, ***p<0.01.*

Table A11: Structure Quality Balance in High- and Low-Reputation Match-ups

Matchup subsample	Loser structure quality (1)	Winner structure quality (2)	Difference: (lose - win) (3)	Std. error of difference (4)	Observations (5)
Panel A. Resolution (Å)					
High scoops High	2.566	2.496	0.070	(0.202)	724
Low scoops Low	2.362	2.258	0.104	(0.123)	491
High scoops Low	2.193	2.183	0.009	(0.058)	520
Low scoops High	2.148	2.155	-0.007	(0.050)	697
Panel B. R-free goodness-of-fit					
High scoops High	0.256	0.250	0.006	(0.004) *	701
Low scoops Low	0.246	0.243	0.003	(0.004)	486
High scoops Low	0.242	0.245	-0.003	(0.004)	512
Low scoops High	0.240	0.238	0.002	(0.004)	695

Notes: This table compares structure quality metrics of winning and losing projects in subsamples of races divided by team reputation as measured by predicted citations. Lower values of resolution and r-free represent better quality. Observations are at the structure level. Column 1 shows the means of the losing projects in the racing sample, and column 2 shows the means of the winning projects in the racing sample. Column 3 shows the difference between the losing and winning projects, and column 4 shows the heteroskedasticity-robust standard error of the difference.
**p<0.1, **p<0.05, ***p<0.01.*

Supplemental Material (not copyedited or formatted) for: Ryan Hill, Carolyn Stein. 2025. "Scooped! Estimating Rewards for Priority in Science."
Journal of Political Economy 133(3). DOI: https://doi.org/10.1086/733398.

References

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, 1990, *215* (3), 403–410.

Berman, Helen, Kim Henrick, Haruki Nakamura, and John L Markley, "The Worldwide Protein Data Bank (wwPDB): Ensuring a Single, Uniform Archive of PDB Data," *Nucleic Acids Research*, 2006, *35*, D301–D303.

Brown, Eric N. and S. Ramaswamy, "Quality of Protein Crystal Structures," *Acta Crystallographica Section D*, 2007, *63*, 941–950.

Campbell, Philip, "New Policy for Structural Data," *Nature*, July 1998, *394* (6689), 105.

Card, David and Stefano DellaVigna, "What Do Editors Maximize? Evidence from Four Economics Journals," *The Review of Economics and Statistics*, 2020, *102* (1), 195–217.

Dasgupta, Partha and Paul A. David, "Toward a New Economics of Science," *Research Policy*, 1994, *23*, 487–521.

Hill, Ryan and Carolyn Stein, "Race to the Bottom: Competition and Quality in Science," *Working Paper*, 2024.

Kim, Soomi, "Shortcuts to Innovation: The Use of Analogies in Knowledge Production," *Working Paper*, 2023.

Milojević, Staša, "Accuracy of simple, Initials-Based Methods for Author Name Disambiguation," *Journal of Informetrics*, 2013, *7* (4), 767–773.

Torvik, Vetle I. and Neil R. Smalheiser, "Author name disambiguation in MEDLINE," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2009, *3* (3), 11.

_ , Marc Weeber, Don R. Swanson, and Neil R. Smalheiser, "A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation," *Journal of the American Society for Information Science and Technology*, 2005, *56* (2), 140–158.

Zhuo, Ran, "Exploit or Explore? An Empirical Study of Resource Allocation in Scientific Labs," *Working Paper*, 2022.