Citations Systematically Misrepresent the Quality and Impact of Research Articles: Survey and Experimental Evidence from Thousands of Citers

Misha Teplitskiy1,5*, Eamon Duede2,6, Michael Menietti3,5, Karim R. Lakhani3,4,5

Affiliations:
1University of Michigan
2University of Chicago
3Harvard University
4National Bureau of Economic Research
5Laboratory for Innovation Science at Harvard
6Knowledge Lab

* Correspondence to: Misha Teplitskiy, tepl@umich.edu. School of Information, 4322 North Quad, 105 S State St, Ann Arbor, MI 48109.

Abstract

Citations are ubiquitous in evaluating research, but how exactly they relate to what they are thought to measure - quality and intellectual impact - is unclear. We investigate the relationships between citations, quality, and impact using a survey with an embedded experiment in which 12,670 authors in 15 academic fields describe ~25K specific referencing decisions. Results suggest that citation counts, when equated with quality and impact, are biased in opposite directions. First, experimentally exposing papers’ actual citation counts during the survey causes respondents to perceive all but the top 10% cited papers as of lower quality. Because perceptions of quality are a key factor in citing decisions, citation counts are likely to endogenously cause more citing of top papers and equating them with quality overestimates the actual quality of those papers. Conversely, 54% of references had either zero or minor influence on authors who cite them, but references to highly cited papers were ~200% more likely to denote substantial impact. Equating citations with impact thus underestimates the impact of highly cited papers. Real citation practices thus reveal that citations are biased measures of quality and impact.


**Introduction**

Citations and metrics derived from them, such as $h$-indices and journal impact factors, permeate all spheres of academic life. The demand for metrics is understandable: the size of the academic literature doubles every 8–9 years (Bornmann & Mutz, 2015), and researchers and administrators seek quantitative and justifiable ways to make choices among unfamiliar projects, people, and institutions. Citations are a particularly attractive metric, because they are widely assumed, or at least wished, to be an unobtrusive, objective, and non-reactive proxy for the quality and intellectual impact of research (Smith, 1981, pp. 84–85). It comes as no surprise then that researchers with more citations are more likely to receive grants, prestigious awards, and desirable jobs (Aksnes et al., 2019; Bornmann & Daniel, 2008; Cronin, 2005, pp. 125–129). Moreover, metrics increasingly inform policy on the grounds that they enable empirically comparing outcomes of policy decisions (Fortunato et al., 2018).

Just as widespread as the reliance on citations is the discomfort with their use (Aksnes & Rip, 2009; DORA – San Francisco Declaration on Research Assessment (DORA), n.d.; Hicks et al., 2015). In addition to statistical and technical issues that limit the reliability of citation metrics (“Not-so-deep impact,” 2005), there are long-standing questions about the psychological and social mechanisms by which they are generated (Bavelas, 1978; Bornmann & Daniel, 2008; MacRoberts & MacRoberts, 1989). Here, we consider two of these questions and assess whether answers to them reveal biases in how citation counts relate to quality and academic impact. First, we investigate whether citations are indeed non-reactive, or whether they actively change the perceptions of scientific quality they aim to measure. Second, we investigate whether authors cite famous and obscure works for different reasons, and whether this distribution of motives biases the relationship between citation counts and influence.

**Citations and quality**

Substantial survey and observational evidence suggests that quality of papers is a key consideration in authors’ citing decisions (Baldi, 1998; Bornmann & Daniel, 2008; P. Wang & Soergel, 1998). Therefore, better perceptions should cause more citations. However, the quality of papers may not be readily apparent (Azoulay et al., 2013) and in a world where citation counts are public and salient, causality may also operate in the reverse. For example, high citation counts could induce high perceived quality, and thus lead to yet higher citation counts. If citations cumulative in such a manner, then equating citations with quality will tend to underestimate (overestimate) the true quality of lightly (highly) cited works.

A number of studies have examined pathways of cumulative advantage in science, famously referred to as the Matthew Effect (Merton, 1968). These pathways include funding, publishing, technological change, and so on. However, the most credible evidence concerns funding, where sharp cut-offs enable quasi-experimental research designs (Bol et al., 2018; Y. Wang et al., 2019). While funding is undoubtedly crucial in many areas of science, it may

---

1 As further evidence consider the infrequency of “negative” citations (Catalini et al., 2015).
Institutions, such as funding agencies and publishers, suggest they prioritize the influential citations. In contrast, direct evidence for cumulative advantage at the paper level, in which a paper’s success begets more success, is missing. In particular, it is unclear whether citations or similar status signals directly shift perceptions of the paper, holding its actual quality constant, although the indirect evidence is highly suggestive (Azoulay et al., 2013). Furthermore, given that citations are routinely visible in high-stakes evaluations like hiring, promotions, and funding, and that many evaluated works are never actually read (Rekdal, 2014; M. Simkin & Roychowdhury, 2006), it is crucial to determine whether citations change perceptions.

Citations and impact

The relationship between citations and impact is seemingly more straightforward. While research can impact audiences beyond academia (Bornmann, 2013), here we focus on the intellectual impact of research on other researchers, as this continues to be the type of impact cherished by many scientists and funders. Therefore, we equate impact with the extent to which a paper changes the theories or research choices of its readers. In practice, impact is treated as nearly synonymous with citations. However, there is copious evidence that many papers are referenced for purposes of persuasion or signaling, rather than to denote intellectual impact upon the author (Bornmann & Daniel, 2008; Cozzens, 1989; Gilbert, 1977; Liu, 1993; Small, 1978). At a high level, authors may cite works that were key in producing their contribution to a literature (“influential”) or works that help them position their contribution within disciplinary contexts (“rhetorical”) (Bornmann & Daniel, 2008; Cozzens, 1989; Jurgens et al., 2018; Nicolaisen, 2007. Previous research has typically approached this dual nature of citations - as a metric generated by normative acknowledgment of influence and as a socially constructed persuasion device (Cozzens, 1989, p. 440) - from one of those directions. For example, scholars have tried to weigh, as precisely as possible, the evidence for each perspective, and adjudicate whether citations can or cannot be meaningfully used to evaluate quality and impact (Baldi, 1998; White, 2004). Here, we pursue an alternative track: we assume that both perspectives have merit and instead investigate whether they introduce systematic bias, rather than completely invalidate, citation metrics. In particular, if some citations are more meaningful than others but their distribution is random, then rhetorical practices add noise but not bias. However, if certain kinds of citations are systematically more meaningful, then there is systematic bias for which one may be able to account. Here, the kinds of citations we consider are those to famous vs. obscure works.

In sum, we pose two questions about citations, quality, and impact.

1. Do citation counts change rather than merely proxy perceptions of quality?
2. Do citation counts accurately measure intellectual impact of famous vs. obscure papers?

---

2 While both citation types are important for effective communication, the mission statements of scientific institutions, such as funding agencies and publishers, suggests they prioritize the influential citations.
Data and methods

Recent efforts have used indirect measures of research impact. One approach assumes that influential works change the language of subsequent works (Gerow et al., 2018). Others rely on third-party annotators to classify references on the impact they had on authors (Jurgens et al., 2018). In this study, we use a direct measure of impact. Using a large-scale survey, we directly ask the authors themselves about the impact of papers they cited in their own work. An embedded experiment allows us to measure the effect of citation information on perceived quality.

We fielded a web-based survey of scientists across 15 fields of science and the humanities, in which we asked about specific references they made in their papers. After verifying the author’s identities and providing IRB information, authors were asked to evaluate the extent to which the paper influenced (impacted) their research choices, as well as the paper’s quality. Perceived quality was elicited by asking, “Rate this reference against others in the field on the following characteristics, with 50th percentile denoting the typical paper.” The characteristics asked of all respondents were overall quality, novelty, significance, validity, generalizability. The survey was relatively short, with median time spent to complete the instrument of 6.97 minutes.

We used rich data from the complete Clarivate Analytics Web of Science (WoS) to systematically sample the literature and survey the scientific community. The Web of Science attributes journals to disciplines, and disciplines to six major subjects - Arts and Humanities, Clinical, Pre-clinical & Health, Engineering & Technology, Life Sciences, Physical Sciences, and Social Sciences. We could not survey all authors in all disciplines, and instead used the following procedure to choose a subset of disciplines. In each subject we sought to identify disciplines with high coverage by WoS and where citation-based metrics were likely to be salient. Accordingly, for each discipline, we averaged impact factors (IF) of its top five journals, and then ranked disciplines according to this average. For each of the six major subjects, we took the 2-3 highest average-IF disciplines. The selected disciplines were biochemistry & molecular biology, physical chemistry, economics, endocrinology & metabolism, energy & fuels, electrical & electronic engineering, history & philosophy of science, immunology, linguistics, nanoscience & nanotechnology, oncology, pharmacology & pharmacy, applied physics, psychology, and telecommunications.

For each discipline, we identified all research articles published in the years 2000, 2005, and 2010 and ranked them according to the number of citations they had accrued through the year 2015. For each percentile of this discipline-year specific distribution, we randomly selected five papers (cited papers), and five random papers citing each of those papers in 2015 (citing papers). If five citing papers were not available, we randomly selected other

---

3 Radichi et al. 2017 pursue an analogous approach, although their work focuses on self- vs. external citations (Radicchi et al., 2017).
4 Protocol 17-1320-01 was approved by the Institutional Review Board of Harvard University.
5 Uncited papers were included in constructing the distribution. Consequently, the lowest percentile for which citers exist is > 1.
papers in that percentile and repeated the procedure. The corresponding author of each citing paper was contacted with a personalized survey (see Supplementary Materials: Materials and Methods for details).

In the survey, we experimentally manipulated the information respondents observed when evaluating papers. The control (85%) and treatment (15%) forms were identical except the treatment form displayed the following status signal, “Our records indicate that this paper has been cited X time(s), which ranks it in the [top/bottom] Y percentile among all papers published in the field in [year of publication],” where X was the paper's true citation count in Web of Science in 2015 and Y was the true percentile in the citation distribution.6

We sent email solicitations to 63,049 corresponding authors of the citing papers. From this risk set, 20.2% (n=12670) of the recipients opened the provided link, and about 15.0% (n=9425) reached the last page.7 Responses came from around the globe: 18.5% from U.S., 12.3% from China, followed by Germany, Japan, and England with 4-5% each. Among respondents, 51.5% reported employment as full, associate, or assistant professor, and 70.3% identified as male. Male authors were more likely to reply, while authors publishing in high impact journals were slightly less likely to reply. See Supplementary Materials: Response rates for further details. We removed self-citations (~9%) from the dataset.8

We measure the effect of providing a signal of citation status on how respondents evaluated cited papers on several dimensions of quality.

If citation counts affect only the number of people who see or read a paper but do not change their perceptions of it, then exposing citation counts should not affect respondents’ ratings. On the other hand, if the treatment changes perceptions, it may have opposite effects when citation counts are relatively low vs. high. High citation counts, indicated by the count and the word “top” in “top Y percentile”, constitute a positive signal that may improve perceptions and vice versa for counts in the “bottom” percentiles.

**Results**

To visualize effects on perceived “overall quality” across the citation distribution, we display in Figure 1 Panel A the control (blue) and treatment (orange) observations, and superimpose a cubic curve for each condition. The control curve shows a small positive relationship between perceived quality and citation counts. However, the treatment curve exaggerates that relationship, and is consistently below the control curve everywhere but for the top few percentiles. To quantify the effects, we examine them for the bottom and top 50% of the citation distribution. We use a linear regression model of the form

---

6 Supplementary Materials: Survey materials displays samples of the two forms.
7 About 10% of emails were undeliverable, so the actual rates of opening and completing are somewhat higher.
8 Self-citations were self-reported.
\[ Y_{ij} = \beta_0 + \beta_1 \text{above.median} + \beta_2 \text{status.signal} + \beta_3 \text{above.median} \times \text{status.signal} + \beta_4 X_{ij} + \varepsilon_{ij} \] (1)

where \( Y_{ij} \) is the rating of a quality attribute by citer \( i \) of paper \( j \), \( \beta_0 \) is the mean attribute for bottom-50\% papers in control condition, \( \text{above.median} \) an indicator variable for top-50\%, \( \text{status.signal} \) an indicator for treatment, \( X_{ij} \) is a set of controls of the citer and paper, and \( \varepsilon_{ij} \) is the error. Errors are clustered by respondents, and the controls consist of the paper’s discipline and publication year, the citer’s gender and position, and whether the respondent or co-author added the reference.

Figure 1 Panel B displays these treatment effects for the five dimensions of quality respondents rated.

![Graph](Image)

**Fig. 1** A. Scatter plot of perceived “overall quality” and citation percentile of Control (blue) and Treatment (orange) observations. Separate cubic curves are superimposed on each group of observations. B. Treatment effects of status signal on five dimensions of perceived quality. The effects are estimated separately for references below the median in discipline-year citation distribution (“Bot. 50\%”) and above it (“Top 50\%”).

The panels show a consistent pattern: exposing citation counts for the bottom half of papers causes perceptions of overall quality, significance, generalizability, and novelty to fall. The effect for perceived validity matches others in direction, but does not reach statistical significance. Meanwhile, exposing citations has no substantial effect on papers in the top half: the point estimate of the effect for top-50\% papers is statistically indistinguishable from 0. *Supplementary Materials: Status signal experiment* presents alternate analyses of these data, which reach qualitatively similar conclusions. Indeed, if treatment effects are modeled as linear, the status signal harms the perceived “overall quality” for the bottom 90\% of papers.

The effect sizes are modest, ranging from 3.4\% (validity) to 19.4\% (significance) of a standard deviation of the respective ratings. Another way to contextualize these effects is in terms of the relationship between perceived quality and citations for the control observations.\(^9\) Here, the results are substantial. For example, a 4.2-point drop in perceived

---

\(^9\) We do not refer to the relationship among control observations as the “ground truth relationship because” the control observations are likely to have also been influenced by status signals outside the present study.
generalizability of a paper makes it about equivalent to a paper that is $4.2 / 0.112 = 37.5$ percentiles lower in the citation distribution.\(^\text{10}\)

Next, we examine the relationship between citations and influence. The influence of a reference was measured directly using the question, “How much did this reference influence the research choices in your paper?” Answer choices ranged from 1 (very minor influence) to 5 (very major influence). The full text of the answer choices and the distribution of responses is displayed in Figure 2.

![Answer choices and counts of responses to the question “How much did this reference influence the research choices in your paper?”](image)

**Fig. 2.** Answer choices and counts of responses to the question “How much did this reference influence the research choices in your paper?”

Authors’ answers reveal that the majority of references (53.9\%) had at most “minor” influence (1 or 2) on their research choices. If the population sample represents the broader population, then most of the roughly 3B references in the WoS do not primarily reflect intellectual influence but, rather, authors’ rhetorical needs. However, references made to the most famous papers reflect substantially more intellectual influence. Figure 3 displays responses across (log) citation counts.

---

\(^{10}\) 0.112 is the slope between generalizability and percentile estimated from control observations. See Table S5 in the Supplementary Materials for details.
Fig. 3. Black circles are a jittered scatter plot of references’ citation counts and perceived influence. Blue curve is a linear regression fit predicting mean influence. The orange curve is a logit regression fit predicting influence $\geq 4$.

Papers with about 1000 citations exert, on average, moderate influence (~3) per citation, while those with 1 citation exert only minor influence (~2). It is worth highlighting that even at the extreme tail of the citation distribution, typical influence is only ‘moderate’.

It is instructive to consider not only averages, but references that exert ‘major’ influence on their readers, i.e. $\geq 4$. These papers alter the guiding theories or questions of their readers and, sometimes, motivate entirely new projects. Accordingly, we define an indicator variable “major_influence” and estimate a logistic regression predicting this outcome from a reference’s citation count, discipline, publication year, and other controls (for details see Supplementary Materials: Influence and citations). Predicted probabilities of major influence are displayed in orange in Figure 3. The curve shows that the probability that a reference had a major influence is low (~10%) for references with 1 citation but about 3 times as high for famous papers (3162 or more citations). Given the high bar for major influence, it is striking that more than a third of famous references meet it.

**Conclusion**

This study yields several key insights concerning the relationship between citations and the reward system of science, all with potentially profound policy implications. First, citation counts are neither an accurate proxy of quality nor of influence. Counts systematically disadvantage the vast majority of scholarly papers by devaluing perceptions of their quality. If perceived quality is a key consideration in citing decisions, biases in perceptions will

---

11 Interestingly, *validity* is the aspect of quality least affected by the citation status signals. In Figure 1 and alternate analyses in the *Supplementary Materials*, we find no statistically significant effects of the signal. We offer two interpretations that necessitate further research. First, validity may be the most objective aspect of science, and citers feel confident in evaluating it without external opinions. A second and more discouraging interpretation is that citers perceive others’ citing decisions to be quite unrelated to papers’ validity, and so those citation decisions do not carry much signal for validity.
induce biases in citing, which in turn bias perceptions even more, and so on. This dynamic of cumulative advantage implies that citations, when equated with quality, overestimate the quality of highly cited papers.

How large that overestimate is depends on the effect size that status signals exert. We suspect the modest effect sizes reported here are an extreme lower bound. Our sample of researchers likely had some broad sense of the citation status of their references at the time of citing. For these respondents, the status signal thus serves as a reminder, rather than revelation. If status effects are strongest when uncertainty is highest (Azoulay et al., 2013; Podolny, 2010) (e.g. when initially searching the literature), then for researchers encountering publications for the first time, citation effects on perception would only be higher. Second, citation counts are but one of the numerous status signals that accompany works as researchers search for and read them. Other ubiquitous signals include authors’ names, affiliations, and publication venues (McKiernan et al., 2019; Tomkins et al., 2017). Consequently, the combined effect from all status signals may, in fact, be very substantial.

At the same time, citation counts underestimate the impact of highly cited papers. Across the citation distribution, the majority of citations in the scholarly literature signal little to no influence. However, citations to highly cited works are systematically more meaningful, with each citation being as much as 200% more likely to substantially affect authors’ research choices. Equating citations with impact thus underweights those meaningful citations, and disadvantages top papers. In short, famous papers influence scientific practice much more than citation counts lead on.

While we directly identify a pathway of cumulative advantage in science - via the causal link between citations and perception of quality - it is unlikely to be the only pathway. For example, previous work has identified cumulative advantage dynamics in grant funding competitions (Bol et al., 2018), and those brought on by awards (Azoulay et al., 2013) and changes in technology (Evans, 2008). Additionally, individuals may invest more effort into reading highly cited works, be more influenced by them, and hence be more likely to cite them (Merton, 1968). These mechanisms would operate in addition to, or interaction with, the effect we identify, and likely amplify it.

This discussion leads to several policy implications. First, our evidence base for the utility of metrics is more limited than previously acknowledged. Metrics are typically defended on the basis of how well they correlate with experts’ subjective assessments of quality (Aksnes et al., 2019; Bornmann & Daniel, 2008). For example, perceived quality of NIH grants predicts their citation impact (Li & Agha, 2015), and similarly with UK’s Research Excellence Framework (Traag & Waltman, 2019). The implicit assumption in such exercises is that the subjective assessments are not themselves caused by citations and other metrics. However, our findings show that subject matter experts change their perceptions of quality in the presence of citation metrics. Thus, positive correlations between citation counts and subjective evaluations are not surprising, and are in fact inevitable because the latter depend on the former.

Second, the findings invite us to consider the effects that common practices in science have on attention and impact. Displaying citation counts, as all major search and discovery
platforms do, will skew perceptions of indexed papers. More troubling still, given that most cited papers are not read carefully (M. V. Simkin & Roychowdhury, 2005) (if at all), the citation count alone can create citation inequality that, over time, develops into substantial impact inequality. The ubiquitous visibility of citation counts is thus bound to make them not a proxy for but a vehicle of attention and impact, thereby exaggerating relatively small quality differences into very different trajectories. A full welfare analysis should of course consider what the scientific community gains from these widely available status signals. From the perspective of consumers of metrics, they probably reduce search costs and helps protect against wasteful decisions in reading, hiring, and funding. However, from the perspective of producers, those most likely to benefit from citation-based decisions are arguably already “well off” scientifically. So if the choice is between penalizing the bottom 90% (and rewarding top 10%) or redistributing resources more equitably, the latter may benefit many and hurt few.

Third, the findings point to concrete, although qualitative, ways in which to adjust metrics to better capture the concepts they intend to measure. High citation counts must be discounted if they are to quantify quality, while they must be inflated if they are to quantify impact.

Lastly, our approach adds a new, powerful method to the toolkit of measuring quality and impact in science: large-scale surveys and survey experiments. The latter enable scholars to reach clean, credible conclusions regarding scientific practices, while the former enable parties primarily interested in accurate measurement to solicit unusually deep and wide information on the impact of their programs, scholars, and institutions.
References

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. *SAGE Open, 9*(1), 2158244019829575. https://doi.org/10.1177/2158244019829575

Aksnes, D. W., & Rip, A. (2009). Researchers’ perceptions of citations. *Research Policy, 38*(6), 895–905. https://doi.org/10.1016/j.respol.2009.02.001

Azoulay, P., Stuart, T., & Wang, Y. (2013). Matthew: Effect or Fable? *Management Science, 60*(1), 92–109. https://doi.org/10.1287/mnsc.2013.1755

Baldi, S. (1998). Normative versus Social Constructivist Processes in the Allocation of Citations: A Network-Analytic Model. *American Sociological Review, 63*(6), 829–846. https://doi.org/10.2307/2657504

Bavelas, J. B. (1978). The social psychology of citations. *Canadian Psychological Review/Psychologie Canadienne, 19*(2), 158–163. https://doi.org/10.1037/h0081472

Bol, T., Vaan, M. de, & Rijt, A. van de. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences, 115*(19), 4887–4890. https://doi.org/10.1073/pnas.1719557115

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80. https://doi.org/10.1108/00220410810844150

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science & Technology, 66*(11), 2215–2222.

Catalini, C., Lacetera, N., & Oettl, A. (2015). The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences, 201502280.* https://doi.org/10.1073/pnas.1502280112

Cronin, B. (2005). *The Hand of Science: Academic Writing and Its Rewards.* Scarecrow Press.

*DORA – San Francisco Declaration on Research Assessment (DORA).* (n.d.). Retrieved December 13, 2019, from https://sfdora.org/

Evans, J. A. (2008). Electronic Publication and the Narrowing of Science and Scholarship. *Science, 321*(5887), 395–399. https://doi.org/10.1126/science.1150473

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A.-L. (2018). Science of science. *Science, 359*(6379). https://doi.org/10.1126/science.aao0185

Gerow, A., Hu, Y., Boyd-Graber, J., Blei, D. M., & Evans, J. A. (2018). Measuring discursive influence across scholarship. *Proceedings of the National Academy of Sciences, 115*(13), 3308–3313. https://doi.org/10.1073/pnas.1719792115

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature News, 520*(7548), 429. https://doi.org/10.1038/520429a

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association for Computational Linguistics, 6,* 391–406.
Li, D., & Agha, L. (2015). Big names or big ideas: Do peer-review panels select the best science proposals? *Science, 348*(6233), 434–438. https://doi.org/10.1126/science.aaa0185

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of Citation Analysis: A Critical Review. *Journal of the American Society for Information Science (1986-1998); New York, 40*(5), 342–349.

McKiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., & Alperin, J. P. (2019). Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *ELife, 8*, e47338. https://doi.org/10.7554/eLife.47338

Merton, R. K. (1968). The Matthew Effect in Science: The Reward and communication system of science. *Science, 199*, 55–63.

Not-so-deep impact. (2005). *Nature, 435*(7045), 1003–1004. https://doi.org/10.1038/4351003b

Podolny, J. M. (2010). *Status Signals: A Sociological Study of Market Competition*. Princeton University Press.

Radicchi, F., Weissman, A., & Bollen, J. (2017). Quantifying perceived impact of scientific publications. *Journal of Informetrics, 11*(3), 704–712. https://doi.org/10.1016/j.joi.2017.05.010

Rekdal, O. B. (2014). Academic urban legends. *Social Studies of Science, 44*(4), 638–654. https://doi.org/10.1177/0306312714535679

Simkin, M., & Roychowdhury, V. (2006). Do you sincerely want to be cited? Or: read before you cite. *Significance, 3*(4), 179–181. https://doi.org/10.1111/j.1740-9713.2006.00202.x

Simkin, M. V., & Roychowdhury, V. P. (2005). Stochastic modeling of citation slips. *Scientometrics, 62*(3), 367–384. https://doi.org/10.1007/s11192-005-0028-2

Smith, L. C. (1981). Citation Analysis. *Library Trends, 30*(1), 83–106.

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences of the United States of America, 114*(48), 12708–12713. https://doi.org/10.1073/pnas.1707323114

Traag, V. A., & Waltman, L. (2019). Systematic analysis of agreement between metrics and peer review in the UK REF. *Palgrave Communications, 5*(1), 29. https://doi.org/10.1057/s41599-019-0233-x

Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project. Study I. Document selection. *Journal of the American Society for Information Science, 49*(2), 115–133. https://doi.org/10.1002/(SICI)1097-4571(199802)49:2<115::AID-ASI3>3.0.CO;2-T

Wang, Y., Jones, B. F., & Wang, D. (2019). Early-career setback and future career impact. *Nature Communications, 10*(1), 1–10. https://doi.org/10.1038/s41467-019-12189-3
Acknowledgments: We thank seminar participants at the University of Michigan School of Information, MIT Sloan School of Management, NESTA Innovation Growth Lab, Harvard Kennedy School, University of Chicago Knowledge Lab, International Conference on Computational Social Science, International Conference on Science and Technology Indicators, International Conference on Scientometrics & Informetrics.

Funding: MacArthur Foundation Research Network on Opening Governance; Schmidt Futures #785, University of Chicago’s BIG Ideas Generator

Author contributions: MT was involved in conceptualization, methodology, data analysis, and writing. ED was involved in conceptualization, data curation, and writing. MM was involved in methodology, data analysis, and writing. KRL was involved in conceptualization, methodology, and writing.

Competing interests: Authors declare no competing interests

Data and materials availability: The nature of the data preclude full anonymization. For example, a citation count > 10,000 can identify specific references (and potentially their citers) with relative ease. However, an anonymized version of a subset of the dataset that is sufficient to reproduce the key findings. This anonymized subset will be made publically available.
Supplementary Materials

Survey materials
Nonresponse analysis
Status signal experiment
Influence and citations
Descriptive statistics of dataset
Figs. S1 to S5
Tables S1 to S8
Survey materials
The following text and figures illustrate the survey flow. Figure S1 displays the (anonymized) recruitment email.

[ Figure S1 about here ]

After clicking on the link, respondents proceeded to confirm that the papers was indeed theirs and read IRB information. Next, they proceeded to a randomized page, the two versions of which are displayed in Figure S999. The control (panel A) and treatment (panel B) versions are identical except that treatment includes the reference’s citation status.

[ Figure S2 about here ]

Next, respondents answered questions about their knowledge of the reference, how much it influenced them, which aspects of their work were influenced (Figure S3). To account for ordering effects in answer choices, respondents were randomized into two forms with identical questions but reversed answer choice order. Form A’s answer choices ranged from smallest/least to biggest/most, while form B had the opposite ordering. Next, respondents rated the reference on various dimensions of quality (Figure S4). Lastly, respondents provided some demographic information.

[ Figure S3 about here ]
[ Figure S4 about here ]

Nonresponse analysis
Disciplines
Response rates were measured by clicks on the personalized survey link. Rates varied substantially across disciplines. The lowest response rate came from oncology (12.9%) and the highest (34.1%) from history and philosophy of science. The number of completed responses and response rates by discipline are displayed in Figure S5.

[ Figure S5 about here ]

Impact factor, gender, and status signal
A response was defined as clicking on the personalized Qualtrics link (opened.survey=True). To investigate response rates by impact factor of the respondent’s journal, gender, and assignment to status signal condition, we estimate a logistic regression with opened.survey as the outcome regressed on the Web of Science 2015 impact factor, a 3-level gender variable [“Unknown”, “Male”, “Female”], and status signal indicator. The model was specified as follows:

\[ \logit(Y_i) = \beta_0 + \beta_1author.female + \beta_2author.male + \beta_3source.impact.factor + \beta_4status.signal + \varepsilon_i \] (1)
Estimates from this regression are displayed in Table S1: Model 1 and average marginal effects are displayed in Table S2: Model 1. Author gender was inferred using Genderize.io (https://genderize.io/). Average marginal effects indicate that authors of papers in high impact journals were somewhat less likely to respond, with response rate decreasing by 0.73% per unit impact factor, while “Female” and “Male” authors were 2.2% and 5.7%, respectively, more likely to respond than “Unknown.” As expected, the randomly assigned status signal did not significantly affect response rates.

Of more concern is whether the status signal affected survey completion rates. Estimates from a logistic regression predicting completed.survey=True are displayed in Table S1: Model 2. The point estimate for the average marginal effect of status.signal on completion is 0.0071 (0.71%), with SE = 0.004 and p = 0.071. Thus we conclude that exposing citation signals increased completion rates by a very small, if any, amount.

**Status signal experiment**

**Average treatment effect**

Status can be high or low, and we expect high status signals to improve perceptions of quality, and low signals to harm them. The expected heterogeneity of the treatment makes the average treatment effect (ATE) a poor summary of it. Nevertheless, we provide the ATE in Table S3 below, estimated with the following regression specifications.

\[
\text{attribute}_{ij} = \beta_0 + \beta_1 \text{status.signal}_i + \beta_2 X_{ij} + \epsilon_{ij} \tag{2}
\]

In this specification, \(\text{attribute}_{ij}\) is the rating of a quality attribute by author \(i\) of reference \(j\), \(\text{status.signal}_i\) is an indicator of whether author \(i\) received a status signal and \(\beta_1\) measures the ATE, and \(X_{ij}\) is a set of control variables describing author \(i\) and reference \(j\). Control variables are indicators for the author’s gender and academic position, and indicators for the reference’s discipline, publication year, whether it was added by the responding author or a co-author, and whether it appeared first or second in the survey.
Most columns (attributes) in Table S3 show the ATE of status.signal to be negative, with effects on perceived validity and generalizability reaching significance.

**Heterogeneous treatment effects**

We now present two ways to examine treatment effect heterogeneity. First, we partition the citation distribution into halves above and below the median. The break at the median is natural because at that point the presentation of the status signal changes qualitatively, from “bottom X% of the citation distribution” to “top X% of the citation distribution.” For this analysis we exclude observations right at the median, where the status signal presentation didn’t include the words “bottom” or “top.” The specification is as in expression (1) in the main text and Table S4 displays estimated coefficients.

For papers below median citations, showing the status signal tends to harm quality perceptions, as indicated by negative and statistically significant coefficients of status.signal. The effect on perceptions of quality is less precisely estimated, but is consistent with the other attributes. For papers above the median, the effect of the status signal is positive for all dimensions, although not consistently statistically significant.

Finally, we model attribute ratings as linear functions of log-citations, with a separate intercept and slope for control and treatments observations. The specification takes the following form:

\[
\text{attribute}_{ij} = \beta_0 + \beta_1 \text{percentile}_j + \beta_2 \text{status.signal}_i + \beta_3 \text{percentile}_j \ast \text{status.signal}_i + \beta_4 X_{ij} + \epsilon_{ij} \tag{3}
\]

where \( X_{ij} \) is the same set of control variables used in (2) above. Estimates from this regression are displayed in Tables S5.

The estimates show a consistent pattern: the status signal exaggerates the underlying (control) relationship. The status signal harms the perception of low-cited papers (coefficients of status.signal are negative and, with the exception of “validity,” statistically significant. However the penalty gets smaller as citation counts, and therefore the nature of the signal, improve - coefficients of the interaction effect are positive and, with the exception of “validity,” statistically significant.

To locate the approximate point in the citation distribution at which the status signal becomes a net positive for perceptions, we locate the percentile at which the control and status-signal lines intersect. For “overall quality” that percentile is \( p^* = 4.579/0.053 = 89.8 \). Thus, for all but approximately top 10% highest cited references, the status signal had a negative effect on perceived quality.
Influence and citations

We investigate the relationship between the citation count of a reference and its influence on the citer’s research choices using mixed models with author fixed effects. Including fixed effects enables us to examine within-person variation in influence, i.e. two references in an author’s reference list. This approach accounts for the possibility that the composition of authors varies significantly across the citation distribution of references. For example, authors citing lowly and highly cited works may have different standards for “influence.” Within-person specifications allow us to examine whether relationships apparent in raw data are due to influence change or composition change. We use the following specification:

\[ influence_{ij} = \alpha_i + \beta_0 \log.cites_j + \beta_1 X_{ij} + \epsilon_{ij} \tag{4} \]

The indices \( i \) and \( j \), enumerate authors and references, respectively. The author fixed effects \( \alpha_i \) capture baseline differences among authors in the influence upon them of their references. The set of controls \( X_{ij} \) is the same as used in (2) above. \( \log.cites \) is base-10 of the citation count. Estimates from regressions of this form are shown in Table S6. Model (2) uses \textit{percentile} instead of Model (1)’s \( \log.cites \), to show that the relationship between citations and influence is robust to rescaling of citations.

[ Table S6 about here ]

Both models show a strong association between influence and citations of a reference when looking within an author’s reference list. Model (1) shows that, for a given author, the influence on her of a reference is 0.143 higher per unit increase in log-cites. Model (2) is consistent with (1), showing a 0.003 increase in influence per percentile increase.
The Laboratory for Innovation Science at Harvard would like to invite you to take part in a quick (5-10 minute) survey about how researchers reference existing work when writing papers. Although citations and related metrics like the h-index are widely used in academia to evaluate research and allocate resources, the referencing decisions on which they are based are poorly understood.

Consequently, we have selected your paper and want to ask you about two specific references within it. Your input will help us and the broader scientific community assess the validity and limitations of existing ways of evaluating and ranking scientific works, and possibly develop superior alternatives.

Follow this link to the Survey:
Take the Survey
https://harvard.ac1.qualtrics.com/SE/?id=...

Many thanks,
Laboratory for Innovation Science at Harvard
lab@harvard.edu

Follow the link to opt out of future emails:
Click here to unsubscribe
Two forms used for the status signal experiment. 85% of randomly assigned respondents saw the control form (Panel A), which does not show any citation information, and 15% saw the treatment form (Panel B), which displays the true citation count and percentile.
Fig. S3.

Reference:

How well do you know this paper?
- Extremely well (know it as well as my own work)
- Very well (familiar with all findings, data & methods, all limitations and critiques)
- Well (familiar with all findings, data & methods, some limitations)
- Slightly well (familiar with all findings, data & methods)
- Not well (only familiar with main findings)

How much did this reference influence the research choices in your paper?
- Very major influence (motivated the entire project)
- Major influence (influenced a core part of paper, e.g. choice of theory or method)
- Moderate influence (influenced an important part of the paper, e.g. additional analysis)
- Minor influence (influenced a small part of paper, e.g. added sentence(s) to Discussion)
- Very minor influence (paper would've been very similar without this reference)
- Not sure

Which aspects of your paper did this reference influence? (Mark all that apply)
- Only minor influence
- Topic
- Theory or conceptualization
- Data
- Methods
- Other: Please explain

Screenshot illustrating questions about the author's knowledge of the reference and its impact on the author. Randomly assigned half of the respondents saw this ordering of answer choices, while another half saw the reverse ordering.
Panel of questions about perceived quality of the reference. The attribute in the last position was randomized to be “Canonical” or “Prominent.” Data from this last position is not included in the present analyses due to its only indirect relationship with quality, but is available from the authors upon request.
Fig. S5

Response counts and response rates by discipline. Each response, if filled out completely, provides data on references.
Table S1.

|                        | (1) opened_survey | (2) completed_survey |
|------------------------|-------------------|----------------------|
| **Intercept**          | -1.444*** (0.0228) | -1.868*** (0.0260)   |
| **author.gender = female** | 0.134*** (0.0270) | 0.135*** (0.0328)   |
| **author.gender = male** | 0.355*** (0.0240) | 0.427*** (0.0271)   |
| **source.impact.factor** | -0.0458*** (0.0050) | -0.0388*** (0.0055) |
| **status.signal**      | -0.0119 (0.0280)  | 0.0560 (0.0310)     |

| Observations | 62550   | 62550   |
| df model     | 4       | 4       |
| pseudo-R     | 0.00518 | 0.00629 |
| LLR p-value  | 2.698e-69 | 9.528e-71 |

Estimates from logistic regressions predicting opening (Model 1) and completing (Model 2) the survey. *p<0.1; **p<0.05; ***p<0.01 for 2-sided Wald tests. Average marginal effects are displayed in Table S2.
Table S2.

|                      | (1)          | (2)          |
|----------------------|--------------|--------------|
|                      | opened_survey| completed_survey|
| author.gender = female| 0.0215***    | 0.0171***    |
|                      | (0.004)      | (0.004)      |
| author.gender = male | 0.0569***    | 0.0541***    |
|                      | (0.004)      | (0.003)      |
| source.impact.factor | -0.0073***   | -0.0049***   |
|                      | (0.001)      | (0.001)      |
| status.signal        | -0.0019      | 0.0071       |
|                      | (0.004)      | (0.004)      |

Average marginal effects from logistic regression tables of Table S1. *p<0.1; **p<0.05; ***p<0.01 for 2-sided Wald tests.
### Table S3.

**Dependent variable:**

| overall.quality | novelty | validity | generalizability | significance |
|-----------------|---------|----------|------------------|--------------|
| (1)             | (2)     | (3)      | (4)              | (5)          |
| status_signal   | -1.255** | -0.889   | -1.381**         | -2.335***    | -0.601       |
|                 | (0.618)  | (0.743)  | (0.686)          | (0.762)      | (0.643)      |

Controls | Y | Y | Y | Y | Y

Observations | 7,430 | 6,894 | 6,910 | 6,759 | 7,406
R^2          | 0.034 | 0.018 | 0.015 | 0.020 | 0.025
Adjusted R^2 | 0.030 | 0.014 | 0.011 | 0.015 | 0.021
Residual Std. Error | 17.250 (df = 7400) | 19.655 (df = 6864) | 17.977 (df = 6880) | 19.768 (df = 6729) | 18.068 (df = 7376)

Estimates from OLS regressions of quality attribute ratings on status signal. Robust standard errors clustered at respondent level. + p<0.1; * p<0.05; ** p<0.01; *** p<0.001 for two-sided t-tests. Constant and controls not shown.
Table S4.

Dependent variable:

|                | overall_quality | novelty | validity | generalizability | significance |
|----------------|-----------------|---------|----------|------------------|--------------|
| (1) above.median | 4.905***         | 4.175***| 3.334*** | 4.194***         | 4.708***     |
|                | (0.546)          | (0.636) | (0.584)  | (0.651)          | (0.573)      |
| status_signal  | -2.802*          | -3.582* | -1.820   | -4.828**         | -3.248*      |
|                | (1.184)          | (1.468) | (1.324)  | (1.459)          | (1.301)      |
| above.median X | 2.164            | 3.782*  | 0.621    | 3.509*           | 3.608*       |
| status.signal  | (1.357)          | (1.641) | (1.531)  | (1.665)          | (1.464)      |

Controls: Y Y Y Y Y Y

Observations | 7,400 | 6,868 | 6,882 | 6,732 | 7,377
R² | 0.049 | 0.029 | 0.021 | 0.030 | 0.040
Adjusted R² | 0.045 | 0.025 | 0.017 | 0.026 | 0.036
Residual Std. Error | 17.105 (df = 7368) | 19.541 (df = 6836) | 17.928 (df = 6850) | 19.667 (df = 6700) | 17.918 (df = 7345)

Estimates from OLS regressions of quality attribute ratings on status signal, examined above and below the median. Robust standard errors clustered at respondent level. + p<0.1; * p<0.05; ** p<0.01; *** p<0.001 for two-sided t-tests. Constant and controls not shown.
Table S5.

| Overall quality (overall_quality) | Novelty (novelty) | Validity (validity) | Generalizability (generalizability) | Significance (significance) |
|----------------------------------|-------------------|--------------------|-------------------------------------|-----------------------------|
|                  (1)               |                  (2)      |                    (3)                |                        (4)     |                               |
| Percentile         | 0.114***         | 0.107***           | 0.087***                         | 0.112***                   | 0.125***                     |
|                    (0.010)       |       (0.011)     |           (0.010)         |                          (0.011)   |                           (0.010) |
| Status signal      | -4.759**         | -5.370**           | -3.046                           | -6.411**                   | -5.403**                     |
|                    (1.718)       |       (2.069)     |           (1.882)         |                          (2.041)   |                           (1.850) |
| Percentile X       | 0.053*           | 0.070*             | 0.026                           | 0.063*                     | 0.073**                      |
| Status signal      | (0.024)          |          (0.028)       |                       (0.026)    |                          (0.028) |
| Controls           |                  |                    |                                   |                           |
| Observations       | 7,430            | 6,894              | 6,910                            | 6,759                      | 7,406                        |
| R^2                | 0.059            | 0.031              | 0.025                            | 0.032                      | 0.048                        |
| Adjusted R^2       | 0.055            | 0.027              | 0.021                            | 0.028                      | 0.044                        |
| Residual Std. Error| 17.026 (df = 7398)| 19.526 (df = 6863)| 17.888 (df = 6879)              | 19.644 (df = 6728)         | 17.854 (df = 7375)            |

Estimates from OLS regressions of quality attribute ratings on status signal and percentile. Robust standard errors clustered at respondent level. + p<0.1; * p<0.05; ** p<0.01; *** p<0.001 for two-sided t-tests. Constant and controls not shown.
|                      | (1)              | (2)              |
|----------------------|-----------------|-----------------|
| **Dependent variable:** |                 |                 |
| influence            |                 |                 |
| log.cites            | 0.143***        |                 |
|                      | (0.043)         |                 |
| percentile           |                 | 0.003**         |
|                      |                 | (0.001)         |
| Controls             | Y               | Y               |
| FE(author)           | Y               | Y               |
| Observations         | 12,946          | 12,946          |
| R²                   | 0.715           | 0.714           |
| Adjusted R²          | 0.315           | 0.314           |
| Residual Std. Error  | 0.879 (df = 5394) | 0.879 (df = 5394) |

Estimates from OLS regressions of influence on impact factor of the citing journal and log-citations or percentile of the reference. Author-fixed effects and controls included. Robust standard errors clustered at respondent level. + p<0.1; * p<0.05; ** p<0.01; *** p<0.001 for two-sided t-tests. Constant and controls not shown.
Table S7.

|                         | count | mean | std  | min  | 25%  | 50%  | 75%  | max  |
|-------------------------|-------|------|------|------|------|------|------|------|
| percentile              | 25476 | 67.69| 25.39| 2.0  | 49.00| 73.00| 90.00| 99.00|
| target.n.cites          | 25476 | 81.83| 795.68| 1.0  | 9.00 | 20.00| 48.00| 77275.00|
| log.cites               | 25476 | 1.32 | 0.59 | 0.0  | 0.95 | 1.30 | 1.68 | 4.89 |
| target.impact.factor    | 24970 | 3.57 | 4.77 | 0.0  | 1.53 | 2.61 | 3.85 | 131.72|
| influence               | 17154 | 2.57 | 1.11 | 1.0  | 2.00 | 2.00 | 3.00 | 5.00 |
| major.influence         | 25476 | 0.14 | 0.35 | 0.0  | 0.00 | 0.00 | 0.00 | 1.00 |
| knowledge               | 17418 | 3.07 | 1.18 | 1.0  | 2.00 | 3.00 | 4.00 | 5.00 |
| overall.quality         | 13177 | 69.41| 17.32| 1.0  | 60.00| 71.00| 81.00| 99.00|
| significance            | 13084 | 70.55| 17.97| 1.0  | 60.00| 72.00| 83.00| 99.00|
| novelty                 | 12272 | 66.71| 19.71| 1.0  | 56.00| 70.00| 80.00| 99.00|
| validity                | 12358 | 71.78| 17.79| 1.0  | 61.00| 75.00| 85.00| 99.00|
| generalizability        | 11961 | 65.71| 19.74| 1.0  | 54.00| 69.00| 80.00| 99.00|
| self.citation           | 25476 | 0.07 | 0.26 | 0.0  | 0.00 | 0.00 | 0.00 | 1.00 |
| gender                  | 18130 | 1.30 | 0.47 | 1.0  | 1.00 | 1.00 | 2.00 | 3.00 |
| status.signal           | 25476 | 0.15 | 0.36 | 0.0  | 0.00 | 0.00 | 0.00 | 1.00 |
| first.paper             | 25476 | 0.50 | 0.50 | 0.0  | 0.00 | 0.50 | 1.00 | 1.00 |
| added.by.coauthor       | 21318 | 1.01 | 0.42 | 0.0  | 1.00 | 1.00 | 1.00 | 2.00 |

Descriptive statistics of quantitative variables used in the dataset. Qualitative variables like discipline and position (professional rank) not included.
Table S8.

Correlation table. Qualitative variables like *discipline* and *position* (professional rank) not included.