DEMS Department of Economics, Management and Statistics

# DEMS WORKING PAPER SERIES

## Cover Effects on Citations Uncovered: Evidence from *Nature*

Pietro Battiston, Pier Luigi Sacco and Luca Stanca

No. 420 – October 2019

Department of Economics, Management and Statistics
University of Milano – Bicocca
Piazza Ateneo Nuovo 1 – 2016 Milan, Italy
http://dems.unimib.it/

# Cover Effects on Citations Uncovered: Evidence from *Nature*

Pietro Battiston[*], Pier Luigi Sacco[†], Luca Stanca[‡]

October 22, 2019

**Abstract**

Despite the prominent role played by bibliometric indicators for evaluating research, progress in pinning down the determinants of citation flows has so far been hindered by endogeneity issues. Based on 30 years of bibliometric data, we exploit a Regression Discontinuity Design to causally identify the effects that an article featured on the cover of the journal *Nature* has on citations to all articles by its authors. We confirm that, over time, cover articles are cited significantly more than non-cover articles, with this difference being long-lasting. However, when considering all articles by *Nature* authors, we find evidence of a crowding-out effect: the publication of a cover article causes citations to previous articles by its authors to decline sharply relative to citations to articles by non-cover authors.

**Keywords**: Bibliometric indicators, Citation flows, Research evaluation, Cover article.
**JEL Classification**: I23, Z1, C5, O3.

[*]Department of Economics, University of Milano Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milan, Italy. E-mail: me@pietrobattiston.it

[†]Corresponding author. Department of Humanities, IULM University, Via Carlo Bo 1, 20143 Milan, Italy; Bruno Kessler Foundation, Via Sommarive 18, 38123 Povo (TN), Italy; Berkman-Klein Center for Internet and Society and metaLAB (at) Harvard, Harvard University, 23 Everett St # 2, Cambridge MA 02138 USA. Email: pierluigi.sacco@iulm.it; pierluigi_sacco@fas.harvard.edu. Pier Luigi Sacco thanks Federico Capasso for an inspiring conversation on the topic of this article.

[‡]Department of Economics and NeuroMI, University of Milano Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milan, Italy. E-mail: luca.stanca@unimib.it

# 1 Introduction

In the last two decades, the availability of large citational databases has had a profound impact on how research is evaluated, financed and even discussed. Impact is the most often mentioned feature of a research product, mainly because it is easily measurable, and impact metrics can be a key driver for researchers to decide to which journal to submit their own articles. Journals therefore maintain a strong focus on keeping their impact metrics high, and especially so relative to competing journals in the same field. Such a strong emphasis on impact reflects the difficulty of assessing the quality of research, as very large numbers of articles are published daily in increasingly differentiated and specialized fields. Quality is an elusive notion that is neither easily defined nor measured.

Indicators of impact are commonly used in the scientific profession for several purposes, such as tenure decisions [1], the assessment of research performance [2], the evaluation of scientific sources [3], and even the choice of issues to investigate [4]. Each citation that a scientific article receives indicates that the citing author(s) considered it worth referring to. Although this does not necessarily reflect a positive assessment of the article's quality, impact has become a commonly used proxy for quality. Despite this practice being highly controversial [5, 6, 7, 8], the lack of viable, practical alternatives favors its adoption and diffusion. There is widespread awareness that impact itself is a multidimensional notion that is not easily captured by a single indicator, and that different measures may better serve different purposes [9]. However, citation-based indices remain a core focus throughout the scientific profession.

Despite the abundance of citational data, relatively little progress has been made in understanding the determinants of citation flows. The relevant literature is fragmentary and not commensurate in volume to the practical relevance of the issue (e.g., 10, 11, 12, 13, 14). Since other aspects of research (such as originality and methodological rigour) are comparatively difficult to measure, understanding whether impact reflects some *intrinsic* quality of research would also provide insights about the extent to which citations can be considered an appropriate measure of quality of research.

Several factors can affect citation flows, including reputational assets such as scientific prizes and awards, the recognition of a journal by a highly reputable scientific society, or the prestige of the academic affiliation of the author. Among the relatively few authors to identify a causal relation between non-intrinsic characteristics of research and citational success, [15] show that citations to articles by medical researchers increase suddenly after their authors are appointed as *Howard Hughes Medical Investigator*. On the other hand, awards may make scientists more self-critical about their own scientific production. [16] find for instance a *decrease* in productivity for mathematicians who are awarded the Fields medal. It cannot be ruled out that awards even play a demotivating role on subsequent productivity and, as a consequence, impact [17]. Publishing in journals promoted by reputable scientific societies may instead lead to a retrospective boosting of a scholar's scientific reputation, increasing visibility and relevance among peers. [18] shows that citations to articles in the journals of the American Physical Society rise when an author publishes further articles in such journals. [19] shows that, in the economics field, authors' affiliation to a small number of elite institutions ensures wider recognition of their articles.

In this article we examine an additional type of prestige-related influence on citational flows: being featured on the *cover* of a high-impact scientific journal. Our analysis focuses on the journal *Nature*, given its prestige in the scientific community and the existence of previous related studies that examine this journal. [20] show, based on total citation counts from the *Web of Science* database, that articles featured on the cover of *Nature* between 2008 and 2010 obtain on average more citations than articles published on *Nature* in the same period but not featured on the cover. However, the interpretation of this finding is far from obvious. Being displayed on the cover of *Nature* can increase the scientific salience and hence the citability of an article, but it is also possible that a more relevant, or highly citable, article has a higher probability of being selected for the cover. At the same time, an article can be featured on the journal cover not only for its scientific value, but also for factors such as the appeal of the topic, its current relevance for mainstream media, or simply the aesthetic quality of the associated images. The citability of cover articles is therefore a complex topic that deserves deeper investigation, also in view of the scarce attention received so far.

The present work aims at shedding light on this issue by examining the dynamics of "cover effects" over a long time range and, most importantly, by taking into account the entire citational life of the authors of articles published on the journal *Nature*. The paper is structured as follows. Sections 2 and 3 describe the data set and methods, respectively. Section 4 presents the results. Section 5 concludes.

## 2 Data

Our data set was constructed in two steps. First, we considered all research articles published on *Nature* between 1987 and 2016, with detailed information about the content of the cover of each issue,[1] by combining online *Nature* archives, bibliometric information retrieved from the database *Scopus*,[2] and the manual investigation of hardcopies. Since only part of each issue of *Nature* contains original research articles, with other sections being devoted to scientific news or commentaries, we restricted the analysis to documents appearing in the "Articles" section.[3] The resulting dataset (*Nature* sample) covers 1,527 weekly issues of the journal that include 2,443 research articles. For each article, we observe citation flows over a period of up to 30 years (from the year of publication to 2017), for a total of 35,401 article/year combinations. We excluded 6 articles for which information on authors was unavailable.

Second, we collected from *Scopus* citations to all articles published by authors of the *Nature* articles mentioned above. This allows us to analyze the entire citational records of the authors. We excluded *Nature* articles with more than 8 authors (representing 31% of the sample). This restriction was applied for a number of reasons. First, it makes it more plausible that individual publications influence authors' prestige. Second, it reduces heterogeneity in authors' publishing patterns (large collaborations typically produce large numbers of scientific articles). Third, it ensures that the sample size remains tractable. Descriptive statistics for the resulting sample of 487,993 scientific articles (extended sample) are presented in Table 1.

It is worth observing that a given article in the extended sample can be related to more than one *Nature* article: when this is the case, it appears more than once in our database, and hence it has a larger weight in the analysis. We think this is the most appropriate way to take into account such multiplicity. Consider two researchers who co-authored an article on a given journal in 2000, and then individually published one article each on *Nature* in 2003 and 2006, respectively. When studying citations to the 2000

[1] [20], who emphasize the aesthetic aspects of the cover, state that "Nature*'s cover images first appeared in 2001"* – in fact, (colored) cover images were present during the entire period of time analyzed in the present study.

[2] We used the *Scopus* API at `https://dev.elsevier.com`.

[3] While the name of this section underwent slight changes along the years, e.g., "Research Articles", it can be clearly identified during the entire period of interest.

Table 1: Descriptive statistics

| | *Nature* sample Overall | *Nature* sample ≤ 8 authors | Extended sample ≤ 8 authors |
|---|---|---|---|
| Articles | 2,443 | 1,675 | 487,993 |
| Article/year combinations | 35,401 | 26,838 | 6,930,081 |
| Total citations | 931,963 | 541,315 | 39,489,857 |
| Max citations | 10,879 | 10,879 | 32,935 |
| Max citations per year | 1,192 | 1,192 | 3,551 |
| Cover articles | 232 | 140 | 43,600 |

*Note:* data sources are *Nature* archives and *Scopus*. The second column describes the restricted sample of *Nature* papers (with no more than 8 authors) on which the extended sample (third column) is based. In the third column, "Cover articles" are those whose corresponding *Nature* article was featured on the cover.

article, in order to consider the effects of both the 2003 and 2006 *Nature* publications, we consider the 2000 article as two separate observations.

# 3 Methods

We start by considering a regression model aimed at capturing the key features of our citation data. Namely, we consider the following specification:

$$c_{i,y} = \alpha_0 + \alpha_1 F_i + \alpha_2 p_i + \alpha_3 y + \epsilon_{i,y} \tag{1}$$

where $c_{i,y}$ is the number of citations that article $i$ receives in year $y$, $p_i$ is the year of publication, and $F_i$ is a dummy variable taking value 1 if the article published on *Nature* by the authors of article $i$ was featured on the cover.

While the comparison of citations to cover articles with those to non-cover articles is informative, it cannot *per se* provide evidence of a causal cover effect. Articles are not randomly selected to appear on the cover. On the contrary, it can be argued that the selection is related to the quality of the article, which in turn is expected to be positively related to citation flows. Hence, a simple comparison of citation flows between cover and non-cover articles does not allow to disentangle "cover effects" from selection based on (unobserved) quality. We tackle this ambiguity by analyzing the dynamics of citations within a Regression Discontinuity Design (RDD).

We test the presence of a dynamic reputational cover effect. Namely, we hypothesize that if an author is featured on the cover of a high-impact scientific journal, this has a positive effect not only on citations to that article, but also on citations to past and future articles by that author. In order to test this hypothesis, we consider the following empirical specification:

$$c_{i,t} = \beta_0 + \sum_{\tau=-T}^{T} Y_\tau(\gamma_\tau + \beta_{1,\tau} F_i) + \beta_2 p_i + \beta_3 y + \epsilon_{i,t} \tag{2}$$

where $c_{i,t}$ denotes the yearly flow of citations that article $i$ receives $t$ years before/after publication of the relevant *Nature* article (e.g., if the authors of article $i$ published an article in *Nature* in 1998, $c_{i,2}$ is the number of citations to article $i$ in 2000), $Y_\tau$ is a set of relative-time fixed effects, i.e., dummy variables equal to 1 when $t = \tau$. Year of publication ($p_i$) and year of citations ($y$) are defined in absolute terms as in Equation (1), with $\beta_2$ and $\beta_3$ capturing overall time trends in citations. Notice that $y$ refers to *absolute* years (between 1970 and 2017), and $t$ refers to years *relative* to publication of the relevant *Nature* article (from $-T$ to $T$).

The key coefficients of interest are $\beta_{1,\tau}$, capturing the interaction between the cover dummy variable and years since/to publication of the *Nature* article. These coefficients allow us to formulate two main hypotheses. The first is that authors of more cited articles are more likely to be featured on the cover of a *Nature* issue:

$$\beta_{1,\tau} > 0 \text{ for } \tau < 0. \tag{H1}$$

The second hypothesis is that an article featured on the cover of the journal *Nature* has a positive effect on subsequent citations to articles by its authors:

$$\beta_{1,\tau} > \beta_{1,0} \text{ for } \tau > 0. \tag{H2}$$

Equation (2) will be first estimated in the *Nature* sample, as in [20]. This restricted data set, however, does not provide a feasible set up for an RDD interpretation, given that, for *Nature* articles, we do not have citations prior to the publication date of the articles themselves. In order to test hypotheses

H1 and H2, we will then estimate Equation (2) in the extended sample. This allows us to observe citations flows before and after the publication of the relevant *Nature* article, providing an appropriate setup for an RDD analysis and, hence, a causal interpretation of cover effects. Such causal evidence cannot be obtained by looking only at articles published in *Nature*, as the choice of articles to be published on the cover might be non-random. In addition, the extended data set allows us to investigate cover effects on citations to the entire set of publications of a researcher. Finally, by analyzing "pre-*Nature*" citations flows, we are able to assess the hypothesis that cover authors are *ex ante* different from non-cover authors.

# 4 Results

Table 2 reports OLS coefficients for Equation (1) estimated in the *Nature* sample. As expected, we find a positive and significant coefficient for the cover dummy variable: a *Nature* article featured on the journal's cover receives on average about 16 more citations per year than a non-cover article. Consistent with the literature [21], more recent articles receive significantly more citations than less recent articles ($\hat{\alpha}_2$=1.389), while the number of citations to an article is negatively related to year of citations ($\hat{\alpha}_3$=-0.572), as in [22] and [13]. Coefficient estimates are virtually unchanged when controlling for previous citations to other articles of the same authors (column 2).

Figure 1 displays the coefficients obtained by estimating Equation (2) in the *Nature* sample (note that here $\tau > 0$). The left panel shows that, as it is generally found in the literature, the number of citations to an article peaks 3 to 5 years after its publication and then gradually falls over time. The right panel shows that cover effects are positive and significant, reach a peak 2 to 3 years after publication, and are long-lasting: cover articles receive significantly more citations than non-cover articles up to 17 years after publication.

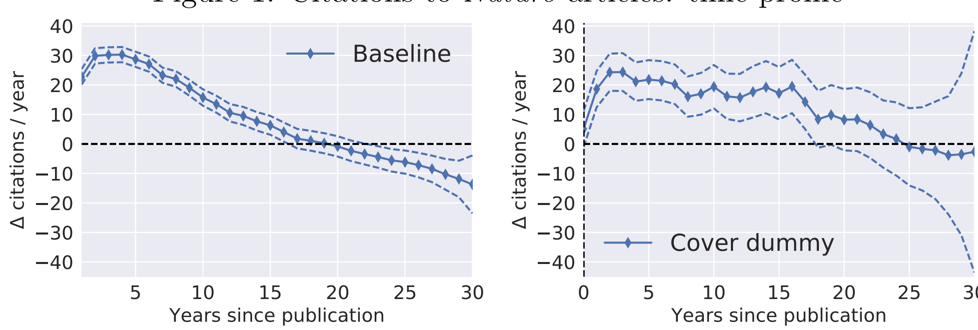
The results presented so far provide clear evidence of a positive difference in citations between cover and non-cover *Nature* articles, consistent with the existing literature [20]. However, they do not allow us to distinguish between the causal effect of being featured on the cover, and a selection effect (more cited authors are more likely candidates for the *Nature* cover). Hence, we turn to the analysis of the extended sample. Table 3 reports OLS estimates

Table 2: Determinants of citations to *Nature* articles

| | (1) | (2) |
|---|---|---|
| Cover dummy ($F_i$) | 16.087*** | 16.050*** |
| | (0.794) | (0.795) |
| Year of publication ($p_i$) | 1.389*** | 1.397*** |
| | (0.034) | (0.035) |
| Year of citation ($y$) | -0.572*** | -0.572*** |
| | (0.038) | (0.038) |
| Previous citations | | -0.000 |
| | | (0.000) |
| $N$ | 35074 | 35074 |

*Note:* OLS estimates for Equation (2) in the *Nature* sample. "Previous citations" refers to citations received by authors of the *Nature* article in the year before its publication.

Figure 1: Citations to *Nature* articles: time profile




*Note:* OLS estimates for Equation (2), sample restricted to *Nature* articles: $Y_\tau$ (left) and $\beta_{1,\tau}$ (right); all articles, all years. N=383,430. Reference category: baseline citations in year 0.

for Equation (1).[4] Similarly to the *Nature* sample, more recent articles receive more citations than earlier articles, while the number of citations to articles by *Nature* authors falls over time. More importantly, the number of citations to articles by "cover authors" is significantly higher than for non-cover authors, when considering either all articles or articles published before the *Nature* article (columns 1 and 2). This is consistent with hypothesis H2. The findings in column 2 are also consistent with hypothesis H1: articles by cover authors *ex ante* receive more citations than articles by non-cover authors. When considering articles published after the *Nature* article (column 3), the number of citations to cover authors is instead significantly lower than for non-cover authors: this finding is at odds with hypothesis H2.

Table 3: Determinants of citations to authors of *Nature* articles

| | All articles | Before | After |
|---|---|---|---|
| Cover dummy ($F_i$) | 0.563*** | 1.328*** | -0.854*** |
| | (0.029) | (0.036) | (0.053) |
| Year of publication ($p_i$) | 0.171*** | 0.136*** | 0.152*** |
| | (0.001) | (0.001) | (0.003) |
| Year of citations ($y$) | -0.130*** | -0.114*** | -0.180*** |
| | (0.001) | (0.001) | (0.003) |
| $N$ | 1146345 | 872813 | 273532 |

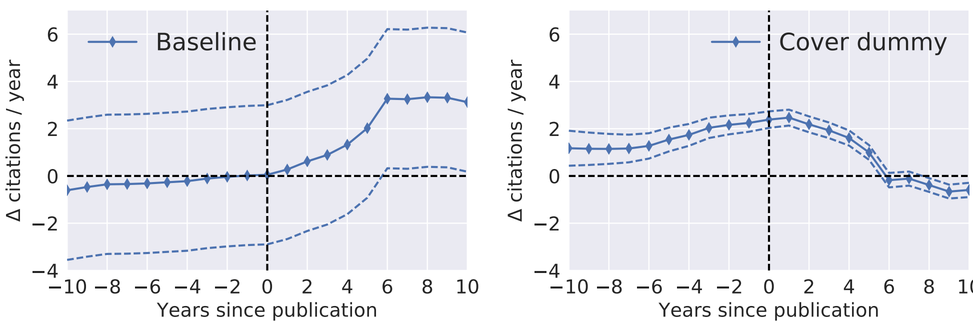
*Note:* the last two columns report estimates based on articles published before and after the *Nature* article. Articles published in the same year are included in the "Before" sample.

Estimates for Equation (2) based on the extended sample (Figure 2) allow us to shed light on these issues. The number of citations to non-cover authors (left panel) is significantly higher following the publication of a *Nature* article. Interestingly, cover authors receive significantly more citations than non-cover authors even *before* publishing in *Nature* (right panel). Specifically, the estimated $\beta_{1,\tau}$ are significantly different from 0 for $\tau \in \{-10 \ldots 0\}$, providing support for hypothesis H1. The year of publication of the *Nature* article marks a discontinuity in citations to cover authors: while the estimated cover

[4] Due to the very large size of the data set, estimation is obtained by taking mean citation flows for all articles related to a given *Nature* article and published in a given year; we then run a weighted OLS on these clusters, where the weights are determined by the number of articles in each cluster.

differential in the number of citations rises over time until publication of the *Nature* article, it falls over time thereafter. In particular, $\beta_{1,\tau} < \beta_{1,1}$, for all $\tau > 1$, with the difference being significant for $\tau > 2$. These findings are at odds with hypothesis H2: cover authors are found to be *penalized* relative to non-cover authors in terms of citations. As a consequence, no significant differences in citations between cover and non-cover authors are found after about 6 years from publication of a *Nature* article.

Figure 2: Cover effects on *Nature* authors: all articles




*Note:* estimates from Equation (2): $Y_\tau$ (top left) and $\beta_{1,\tau}$ (top right); all articles, all years ($T = 10$).

In order to interpret this finding, we consider estimates of Equation (2) for different sub-samples. Figures 3 and 4 focus on articles published before and after the *Nature* article, respectively. Citations flows to articles published before the *Nature* article are virtually identical to those for the entire sample. Instead, the cover effect is small and not significant for articles published after the *Nature* article. This indicates that the discontinuity observed for the entire sample is mainly explained by the change in citation flows to articles published before the *Nature* article. *Ceteris paribus*, publication of a cover article *crowds out* citations to previous articles by its authors.[5]

[5]It should be noted that publishing a cover article results in a reduction of citations to other articles *only* when compared to non-cover authors. The overall effect on citations to articles by cover authors is positive, although not significant.

Figure 3: Cover effects on *Nature* authors: before *Nature* article

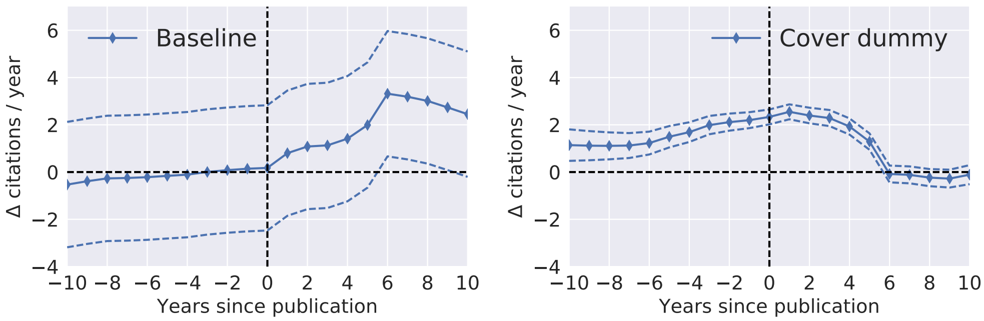
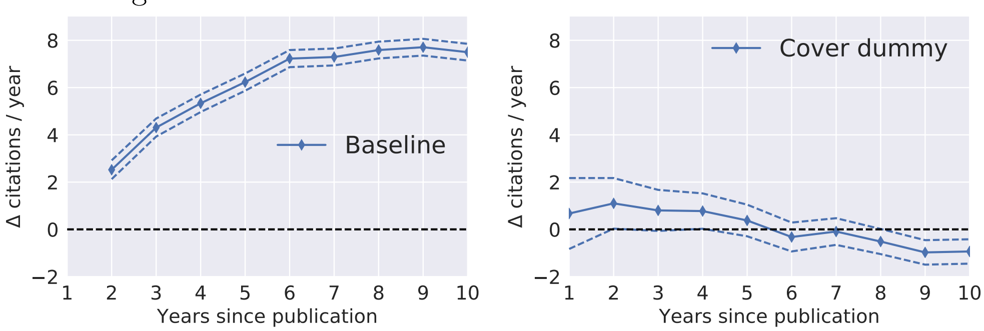



Figure 4: Cover effects on *Nature* authors: after *Nature* article




*Note:* Reference category: baseline citations in year 1.

When considering the left and right plots of Figure 2 together, a complementary interpretation can be suggested: the *ex-ante* relative advantage of cover authors could be offset by the fact that non-cover authors also published an article in *Nature*. For instance, it might be the case that only high-impact researchers are featured on the cover of *Nature* (hence the significance of the cover dummy before year 0), but publishing an article in *Nature* (be it on the cover or not) is one way to become a high-impact researcher – and hence close the gap in citations relative to those who already are.

# 5 Conclusions

We analyze the bibliographic effects of being featured on a journal's cover, by focusing on authors who have published on the journal *Nature*. When considering only articles published in *Nature*, we find that cover articles receive significantly more citations than non-cover articles, consistent with the existing literature, and that this difference is long-lasting. This remains true when controlling for the number of citations obtained by the authors before publication of the *Nature* article.

When considering citations to all the articles by authors who have published in *Nature*, we find that authors whose articles have been featured on the cover receive *ex ante* more citations than non-cover authors. More importantly, following the publication of a cover article, this difference tends to shrink and is short-lasting: the difference in citations between cover and non-cover authors is positive and statistically significant only within five years since publication of the *Nature* article. Being featured on the cover therefore reflects pre-existing differences in scholarly impact, but publishing in a high-impact journal, whether on the cover or not, gradually eliminates this gap. This is due to the fact that a cover article crowds out citations to previous articles by the same author.

In conclusion, is it desirable to be featured on the cover of a high-impact journal? The answer is yes: a cover article receives for several years a higher number of citations relative to a non-cover article. If this may crowd out previous work, it is due to a high relevance and visibility of the novel research. Our results concerning citations to other works are compatible with the possibility that being featured on the cover may be psychologically regarded by the author as akin to receiving a prize. As discussed in the literature, this could also have the effect of making the author more self-critical and vigilant as to the publication of new articles, or even demotivate the author to do further work. Such factors could therefore also play a role in the ex-post equalization of citation flows of cover and non-cover authors. Our analysis does not allow us to give a clear response concerning such channels. The issue is however worth of investigation in future research.

# References


[1] Heckman, J. J. & Moktan, S. Publishing and Promotion in Economics: The Tyranny of the Top Five. Tech. Rep. 25093, National Bureau of Economic Research (2018).

[2] Bordons, M., Fernández, M. & Gómez, I. Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics* **53**, 195–206 (2002).

[3] Brown, H. How impact factors changed medical publishing – and science. *BMJ* **334**, 561–564 (2007).

[4] Monastersky, R. The Number That's Devouring Science. *Chronicle of Higher Education* **52**, 14 (2005).

[5] PLoS Medicine Editors. The impact factor game. *PloS Medicine* e291 (2006).

[6] Simons, K. The Misused Impact Factor. *Science* **322**, 165–165 (2008).

[7] Archambault, É. & Larivière, V. History of the journal impact factor: Contingencies and consequences. *Scientometrics* **79**, 635–649 (2009).

[8] Callaway, E. Beat it, impact factor! Publishing elite turns against controversial metric. *Nature News* **535**, 210 (2016).

[9] Bollen, J., Van de Sompel, H., Hagberg, A. & Chute, R. A Principal Component Analysis of 39 Scientific Impact Measures. *PloS one* **4**, e6022 (2009).

[10] Price, D. d. S. A general theory of bibliometric and other cumulative advantage processes. *Journal of the Association for Information Science and Technology* **27**, 292–306 (1976).

[11] Johnson, D. Getting Noticed in Economics: The Determinants of Academic Citations. *The American Economist* **41**, 43–52 (1997).

[12] Ayres, I. & Vars, F. E. Determinants of Citations to Articles in Elite Law Reviews. *The Journal of Legal Studies* **29**, 427–450 (2000).

[13] Mingers, J. Exploring the dynamics of journal citations: Modelling with s-curves. *Journal of the Operational Research Society* **59**, 1013–1025 (2008).

[14] Di Vaio, G., Waldenström, D. & Weisdorf, J. Citation success: Evidence from economic history journal publications. *Explorations in Economic History* **49**, 92–104 (2012).

[15] Azoulay, P., Stuart, T. & Wang, Y. Matthew: Effect or Fable? *Management Science* **60**, 92–109 (2014).

[16] Borjas, G. J. & Doran, K. B. Prizes and productivity: How winning the fields medal affects scientific output. *Journal of Human Resources* **50**, 728–758 (2015).

[17] Robinson, C. D., Gallus, J., Lee, M. G. & Rogers, T. The demotivating effect (and unintended message) of awards. *Organizational Behavior and Human Decision Processes* (2019).

[18] Battiston, P. Citations are Forever: Modeling Constrained Network Formation. Working Paper 2014/19, LEM Paper Series, Sant'Anna School of Advanced Studies (2014).

[19] Medoff, M. H. Evidence of a Harvard and Chicago Matthew effect. *Journal of Economic Methodology* **13**, 485–506 (2006).

[20] Wang, G., Gregory, J., Cheng, X. & Yao, Y. Cover stories: An emerging aesthetic of prestige science. *Public Understanding of Science* **26**, 925–936 (2017).

[21] Althouse, B. M., West, J. D., Bergstrom, C. T. & Bergstrom, T. Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology* **60**, 27–34 (2009).

[22] Aksnes, D. W. Characteristics of highly cited papers. *Research evaluation* **12**, 159–170 (2003).