Andrew Dillon
School of Information
University of Texas at Austin
Texas, USA.

OUTSOURCING OUR JUDGEMENTS: THE TROUBLE WITH METRICS FOR EVALUATING FACULTY

Abstract

Faculty evaluations are a routine part of academic life, and theoretically serve to ensure fair hiring, promotion, and reward practices. In the search for efficiency, many universities are placing increased emphasis on quantitative measures that are flawed and which incorporate a bias toward team science and multi-authored work. The long-term consequences for certain types of research and scholarship in LIS are discussed, and academic leadership in the field is asked to act.

1. Introduction

There is no shortage of publication venues currently for faculty in the broad library and information science fields. Web of Science lists 86 journals under their 'information science and library science' category, and while this number may represent greater inclusion than some LIS scholars may recognize, Google Scholar can still find a 'Top 20' journals for the LIS field which is hardly exhaustive of even a narrow definition of the field. If we include the full range of publication venues chosen by faculty across the broader iSchool domain, the 86 included in WoS is likely a substantial underestimate. For current faculty therefore, there are few obstacles to finding a venue for their work.

Despite the apparent range of journals on offer, it is perhaps no surprise that some of the most cited papers in the LIS field are themselves bibliometric studies (Kharabati-Neshin et al 2021). Measuring and analyzing a field's output, impact and communication patterns is recognized as a core interest of our field. However, while the bibliometric analysis of disciplines and regions is widely accepted as useful, some have argued its use and application in the evaluation of individual scholarly productivity is more problematic has in fact gone too far (see e.g., Blockmans et al, 2014). Should this concern us? I believe so. Though an originating discipline of bibliometric measures, many programs in Library and/or Information Studies seem to have adopted a narrow and stereotypical approach to research assessment that emphasizes publication quantity at the expense of quality. While there are valid reasons for some use of metrics in faculty evaluation, the naïve application of such measures, particularly in a discipline

that recognizes their limitations, has, I believe, induced a laziness in quality assessment that risks negatively impacting the faculty and research quality of the field.

2. The illusion of objectivity for faculty productivity

A major driver for the metrication of faculty productivity assessment is its apparent objectivity. Faculty publish their research in appropriate venues where it is presumably consumed by peers. This reflects the first assumed quality standard of scholarly communication, the peer-review process, a filter that supposedly controls the quality of published work within a field, and reflects the old adage of faculty life, one must publish or perish. Of late, however, publishing alone seems insufficient. Once published, if one's peers deem the work sufficiently worthy, they will likely reference it in their own publications thereby providing the original author with a citation, the subsequent tallying of which provides a supposedly more reliable and valid index of the original scholar's impact. Over time, as faculty continue to publish, they can build up their citation count and its variant synthesis measures (the *h-index*, *i10-index*, etc.) to provide 'evidence' of their apparent scholarly merit beyond a simple list of their publications.

There is an extensive literature on citation behavior in the field (am grateful to one reviewer's suggestion that Wouters (1999) is a foundational read) but this is not a bibliometrics paper in the traditional sense, it is a call for rethinking what quality means when considering scholarly work. For a balanced comparison of citation theories, Bornmann and Daniel (2008) offer a cohesive review. My argument is not that bibliometrics has no place in faculty evaluations but that evaluations of scholarly merit are distorted by the simplistic application of bibliometric scores. While experts in this domain are often aware of the limitations, their reservations are either insufficiently articulated to influence those who would use them, or they are insufficiently codified to enable appropriate correction. This is an existential threat to the values of scholarly work, and we need to face this challenge head-on.

As we generally understand citations, the true value of this type of measurement rests on the dual provisions that faculty only cite work they find interesting or worthy, and that exposure to such worthy scholarship is distributed evenly across the academy. On the surface, these provisions might seem plausibly met but there is ample reason to question this state of affairs. Boorman and Daniel (2008) review several studies which have revealed that citation choices are almost always influenced by a range of factors that are unrelated to quality (e.g., availability, language, perceived journal quality, social networks and more). This we have known for a long time, yet it continues. Recently Beck et al (2018) examined motivations among Human-Computer Interaction (HCI) researchers for citing specific works in their publications and they reported that while building the paper's argument was one appropriate reason, participants acknowledged that citing work by familiar people (advisors, colleagues, and so forth) was common, as was the use of strategic citations in anticipation of how reviewers might receive the work. In case of the latter two motivations, many authors gain new citations to their work for reasons unrelated to its quality but more reflective of personal connection and power dynamics.

Given the enduring effect of such variables, we must ask, what lessons have we really learned when it comes to assessing research quality and impact?

While we might acknowledge that initial access to research ideas and literatures is gained through the filter of one's doctoral advisors and program, and that personal and social factors play a significant role in citation behavior, there are further complications. Contemporary searching and consumption of research is mediated through a host of filters, often built off bibliometric tools which bias exposure toward papers and thus ideas, theories and models that have already achieved higher impact counts (see e.g., Beel and Gipp, 2009). This technological biasing, in concert with the very human influences on choices made by authors, suggests there is little or no provision for self-correction in this type of citation system, the known biases are thus permanently baked in.

Yet there are more problems with assessing quality through citations that only rarely are acknowledged in research assessments. A citation count is, in its purest sense, an indivisible unit. A paper, regardless of authorship, is cited within another paper and that constitutes a one count of relevance or worthiness (assuming the citation is positive). Even if the paper is referenced multiple times within the citing work, this still registers as a count of one for the paper. But the paper is not the defining unit when employed in faculty evaluations. On the authorship side, a paper can serve as much more than one unit. If two or three authors write a paper, each time it is referenced, each of these two or three authors will themselves count this as one unit of citation to themselves. In a world where more than a dozen authors are associated with a paper, a single citation will find itself adding a point to the impact scores of 12 or more authors, none of whom is likely to have done as much work in producing that one paper as a single author. That alone skews the citation distributions and effectively punishes the independent scholar. Attempts to correct for this (e.g., Toi, 2011) have so far failed to be adopted in typical faculty evaluations and I know of no information program where any such correction is formally applied.

Meanwhile, faculty understand too well the importance of the *h*-index on their careers and rewards, consequently it should not surprise us that we have seen the rise of 'citation farms'. These are small groups of scholars who agree to massively cite each other, thereby driving up their indices producing ultimately measures of little more than their ability to rig the system. Ionnadis et al (2019) studied the patterns of self- and small group citation across 100,000 scholars in multiple fields and concluded that while there is disciplinary variability, median self-citation rates are around 12%, but some scholars self-cite at rates as high as 50%. In other words, for some scholars with large *h*-indices, half of their citations are from themselves. Little wonder Ionnadis et al (2019) concluded that as currently employed, citation metrics are "spurious and meaningless" (para 6).

3. Substituting counts for quality

If most of the shortcomings are known, why are we not counter balancing this more formally in faculty evaluations? In a rush to quantify productivity and impact, faculty have

become experts in measuring their individual citation scores, and universities, hungry to demonstrate their value for money, research rankings, or contributions to the world, devour and promote the numbers accordingly. Junior faculty are frequently advised to consider publication venues with at least one eye on its impact potential. The fear here, of course, is that good work might be invisible to others, consequently failing to register its impact and thereby hurting the career development of the scholar. Note, that in such a case, the quality of the paper and the work it reports has not changed, it is only the means of measuring impact that is affected. A thorough reading of the work by a review committee would presumably reconcile the apparent conflict in assessment but how confident are any of us that such a reading will occur? Reading takes effort. Google Scholar makes the gathering and continual monitoring of one's citation record convenient (even if not entirely accurate, see e.g., Booker et al 2013) and the use of what were once bibliometric technicalities is so accessible now that their widespread use in faculty evaluations, promotion and tenure cases, and recruitment has been normalized.

This has bred, I believe, a form of laziness in quality assessments that may not be widely acknowledged. Faculty routinely discuss other scholars' citation 'scores' (a nice example of the gamification of scholarly life) and tenure letters rarely fail to mention the citation rate of a young scholar as if this were a reliable indicator of long-term potential, which ultimately is what is being assessed. Even at the full professor promotion level, when a candidate's length of time on the career track should allow for a more complete evaluation of productivity and strength of contributions, *h*-index and citation counts are employed by evaluators and committees as a shorthand tool and rationale for their decisions. This may well serve a purpose in evaluations where unpopular scholars or those on the receiving end of bias can leverage impact score to argue for their contribution when others chose to ignore them. But where it works poorly is for scholars in less popular areas of enquiry who might be doing excellent and potentially important work, for a smaller audience, or for an audience yet to catch up with innovations. Unless evaluators make the effort to ignore or at least contextualize the scores, and to base their evaluations on the quality of work submitted by actually reading it, there is the potential for significant damage.

I hear the argument made that the scores are rarely the sole criterion, that review committees examine the body of work and discuss the quality of a scholar's ideas, but one wonders how much the discussion and resulting evaluation are influenced by the citation scores and impact measures? My worry is that the scores anchor the evaluation, that review committees rely on them or try to explain them away, and in the end, it is a rare review process that will deem any scholar with a high impact factor as not performing well.

4. How does this impact our field?

If the metrics are convenient, easy to use, and adopted widely, but are lacking in validity, then there are three negative impacts we can anticipate: technocentrism in research,

disincentivizing of lone scholars for team authorship, and the outsourcing of evaluation in pursuit of efficiency.

4,1. Technocentrism

The type of research that is conducted in LIS will over time tend to narrow or, most likely, follow trends in technology (witness the current rush of publications on AI) and seek to find a home in a range of venues created to address that domain, usually scholarly outlets with highly regulated forms of presentation and delivery, limiting amount and type of content to fit a template and set of genre requirements.

4.2 Authorship impact

Lone scholars will find themselves at a significant disadvantage in terms of placement, promotion, and career prospects. While some might argue that this is a function of the world's inevitable move toward team science, a space in which lone scholars have little ability to conduct the studies now required, there is evidence now questioning the real value of large teams (see e.g., Wu et al, 2019). Disruption rarely springs from large groups of scholars and if we assume the important problems are only those that require team research, we risk making the mistake of allowing method to drive our problem selection.

4.3 Review quality and anchors

Perhaps most crucially, as indices and scores are increasingly used as surrogates for quality, the work of evaluating scholarship too easily becomes outsourced in the pursuit of efficiency. When a time-challenged senior figure in the field is asked to evaluate another junior colleague for tenure, at short notice, and with no inducement to comply other than being a good citizen, then it's easy to understand why metrics are so attractive. But if we only use metrics generated by technical systems which embody their own biases, who will actually read the work to form a quality judgement? Furthermore, once obtained and reported, the anchoring effect of a number raises concerns of how independent the review can be.

Naturally, the reliance on citation scores serves as a significant demotivator for scholarship in less trendy areas of the field. In a community where perhaps only a handful of other scholars are pursuing similar lines of work, there are fewer citation-loading venues in which to publish, and fewer opportunities to raise one's profile in comparison to one's peers. Within iSchools, in particular, where interdisciplinary work is supposedly nurtured, we hope there will inevitably be diversity of work and variation in what might be termed 'citability', so one would also hope that appropriate, more nuanced evaluation protocols are in place, but can we be confident they are? Answering this question is less about examining the written guidelines universities love to generate on evaluation processes and tenure or promotion standards, and more about an honest and critical examination of the closed-door practices of reviewers and committees.

In a recent editorial for *Information and Culture*, Dillon (2020, p 203) noted:

> 'The pressures on scholars to publish often and quickly means that work which requires deep thought, consideration, and a respect for history in the broadest sense is relegated in urgency to the demands of output. This in turn manifests itself in a set of publication opportunities that emphasize the novel, the latest, or the trendiest. In such an environment, it is difficult to encourage scholars, particularly junior faculty, to stand back, reflect, and provide considered treatment of the complex interplay of social and cultural dynamics underlying our information world.'

The academic leaders of the field have a significant role to play here. If we are to support intelligent and innovative scholarship then we must sustain its existence over the long-haul by recognizing and promoting the best work, and ensuring evaluation is not outsourced to an indexing service. The use of metrics and alt-metrics is core to the LIS disciplinary skill set but we risk losing a vital perspective on their application if we fail to articulate their limits and to offer alternatives. Where these alternatives are not so efficient or easy to derive, we must, as scholars and reviewers, accept our responsibility to do the necessary work in evaluating research quality, and recognize the effort involved in doing so as part of our role. Perhaps it is time for academics to revisit their regular assessment and evaluation processes to ensure they are appropriately aligned with our academic mission and to recognize the important work involved in conducting balanced and thoughtful reviews of research contributions.

References

Beck, J., Neupane, B. and Carroll, J (2018) A Study of Citation Motivations in HCI Research, *SocArXiv* doi: 10.31235/osf.io/me8zd

Beel, J. and Gipp, B. Google Scholar's Ranking Algorithm: The Impact of Citation Counts (An Empirical Study). In André Flory and Martine Collard, editors, *Proceedings of the 3rd IEEE International Conference on Research Challenges in Information Science* (RCIS'09), pages 439–446, Fez (Morocco), April 2009. IEEE. doi: 10.1109/RCIS.2009.5089308.

Blockmans, W., Engwall, L, and Weaire, D. (eds.) (2014) *Bibliometrics: Use and Abuse in the Review of Research Performance* London: Portland Press

Boeker, M., Vach, W. and Motschall, E. Google Scholar as replacement for systematic literature searches: good relative recall and precision are not enough. *BMC Med Res Methodol* **13,** 131 (2013). doi: 10.1186/1471-2288-13-131

Bornmann, L.; Daniel, H. D. (2008). "What do citation counts measure? A review of studies on citing behavior". *Journal of Documentation*. **64** (1): 45–80.

Dillon, A. (2020). It's All about Information and Culture, *Information & Culture* 55(3), 201-203.

Ioannidis, J., Baas, J., Klavans, R., and Boyack, K. (2019) A standardized citation metrics author database annotated for scientific field. *PLoS Biology*, doi: 10.1371/journal.pbio.3000384

Kharabati-Neshin, M; Yousefi, N; Mirezati, and Saberi, M., "Highly Cited Papers in Library and Information Science Field in the Web of Science from 1983 to 2018: A Bibliometric Study" (2021). *Library Philosophy and Practice* https://digitalcommons.unl.edu/libphilprac/6251

Toi, R. (2011) Credit where credit's due: accounting for co-authorship in citation counts. Scientometrics, 89(1): 291–299. doi: 10.1007/s11192-011-0451-5

Wouters, P. (1999) The Citation Culture. Ph.D. Thesis, University of Amsterdam http://garfield.library.upenn.edu/wouters/wouters.pdf

Wu, L. et al., (2019) "Large teams develop and small teams disrupt science and technology," *Nature*, doi:10.1038/s41586-019-0941-9, 2019.

Van Noorden, R. and Chawla, D.S. (2019. Hundreds of extreme self-citing scientists revealed in new database *Nature* **572**, 578-579 doi:10.1038/d41586-019-02479-7