Heuristics, not plumage: a response to Osterloh and Frey's Discussion Paper on 'Borrowed Plumes'

Author: Dr Steven Wooding

Affiliations: Research Strategy Office (Senior Research Fellow) and Bennett Institute for Public Policy (Affiliated Researcher), University of Cambridge; Jesus College Intellectual Forum (Senior Research Associate), University of Cambridge

Corresponding author: sw131@cam.ac.uk

Corresponding address: Research Strategy Office, Old Schools, Trinity Lane, Cambridge, CB2 1TN

Part of the virtual Special Section on the Discussion Paper by Osterloh and Frey (2020).

Received 3 July 2019; Received in revised form 12 August 2019; Accepted 9 September 2019

Keywords: heuristics, bibliometrics, journal impact factor, normalisation, research policy, research evaluation

At its simplest, ‘How to avoid borrowed plumes in academia’ by Margit Osterloh and Bruno Frey (Osterloh & Frey, 2020) is a paper about how using Journal Impact Factor (JIF) to judge academic papers and their authors is a bad idea; their suggestion of why people nevertheless persist in doing so; and what might be done to stop them. More broadly, it is a paper about heuristics (or rules-of-thumb) – why they are used; when they should not be used; and how to stop them being used in those contexts. I agree with Osterloh and Frey that JIF is a bad heuristic for judging research, but I find their arguments about why it is used and what might be done to stop people using it unconvincing and impractical. In this short Note, I argue that the use of heuristics is inevitable and, if effectively selected, they can improve decision-making. The challenge for an individual is to decide which heuristics are worth using. The policy challenge is to dissuade people from using inappropriate heuristics – and doing this requires good evidence on how and why the heuristics are being used, something that is missing from Osterloh and Frey’s paper.

Heuristics are an unavoidable fact of life. In a world of information, none of us has the time to fully appraise every option in every situation (Tversky & Kahneman, 1974). For example: if we were to fully assess the nutritional, economic, environmental and social pros and cons of our breakfast cereal choice every morning, we would never get anything done. Often heuristics allow us to make up for incomplete knowledge – I have not tried all the breakfast cereals available at the on-line grocery store but I can use the reviews of like-minded shoppers to improve my chances of choosing something I’ll enjoy. Not only is the world full of information but our lives are also full of things to do. As researchers, we have to write grant applications, conduct research and publish our results, amongst a myriad of other things. We have to decide how best to spend each hour, which is better: polishing that fellowship application or starting a new paper? There is not enough time to dive into detail with everything so we inevitably resort to heuristics: for example using heuristics to assess the quality of research.

Different heuristics simplify to different extents. Compared to JIF, simply counting an article’s citations is a somewhat better, but still rather flawed, heuristic for judging the impact and maybe the quality of research reported in the article. I would argue that sensibly normalised citation counts often provide a useful heuristic for helping to judge research, for example, to shortlist papers for reading in depth or allowing peer review efforts to be focussed on borderline cases; although as Yaqub (2020) notes, we still don't have a well developed understanding of citation, and citations were never intended for evaluation. JIF is a heuristic based on a heuristic, simplifying citation counting by assigning a single value to each journal (and all the articles published in it). This value is the mean number of citations attracted by each article in the journal in the two years following publication.[1] Building on citation counting, JIF inherits all the problems of citation counting and then amplifies them in particularly unhelpful ways.

Two important problems of JIF (ignoring the possibilities of gaming[2] - see e.g. Martin, 2016) arise because the distribution of citation numbers across the articles in journals is heavily skewed and is also very broad. This skewed distribution reflects the fact that there are a few papers with a very high number of citations so the distribution has a higher mean than median, with the result that for most papers published in a particular journal the JIF value is higher than their actual number of citations (Lotka, 1926; Lozano, Larivière & Gingras, 2012; Wallace, Larivière & Gingras, 2009). This is the aspect highlighted in the paper by Osterloh and Frey (2020); they argue that researchers cling to JIF because it makes their papers look better by providing a numerically higher heuristic than the number of citations each article actually attracted. However, the second aspect, the breadth of the citation distribution, is also important. It means that JIF is a very poor predictor of the number of citations for any particular paper. A third problem is arises because citation behaviour varies considerably between fields – biochemists, for example, tend to cite a large number of papers, while mathematicians cite relatively few. Hence, to estimate relative quality across different fields it is necessary to normalise, something JIF fails to do.

[1] Strictly speaking, the Journal Impact Factor is the total number of citations earned by all publications in that journal over the previous two years, divided by the number of 'substantive' research articles published in the journal over that 2-year period. In other words, non-substantive papers (e.g. editorials, book reviews etc.) do not count in the denominator of the JIF equation, even though they may earn citations included in the numerator of the equation. This historical anomaly opens up possibilities for less scrupulous journal editors to manipulate the JIF for their journal.

[2] See previous footnote.

So what was JIF designed to do? JIF started out as a heuristic to help librarians select which journals to subscribe to in an age before detailed readership statistics existed (Garfield, 2006; Wouters et al., 2019). Librarians did not need to read all the journals and make a subjective judgement – they could simply look at the Journal Impact Factor to get an indication of which journals in their area were more heavily cited, suggesting which were the most important. That heuristic could suggest a shortlist, which they could look at in more detail or discuss with their faculty. JIF was probably quite a useful heuristic for that purpose, but is it useful for assessing papers and researchers?

Assessing heuristics involves a cost-benefit judgement. How often will it lead you to the right conclusion? How much time and resources will it save you? And will those savings make up for the cost of the times that it fails? Thinking at a system level the questions widen to ones of collective behaviour: is the heuristic being used enough to matter? Does using the heuristic lead to better outcomes overall? And, very importantly, what will people do if they can't use the heuristic – will they fall back on something worse?

So how does JIF stack up as a heuristic? For judging which journals to subscribe to, it probably has some value. But in most other contexts it is pretty terrible; as well as being inaccurate for individual comparisons, it has many unfortunate consequences for collective behaviour, which Yaqub (2020) summarises.

In this context, Osterloh and Frey (2020) make two main suggestions – that JIF is still widely used and that use is increasing; and that JIF is used because it makes people's 'citation' numbers look bigger. These are potentially valuable insights in the fight against JIF, yet Osterloh and Frey provide no systematic empirical evidence to bolster their suggestions, rendering them little more than speculation.

In the first case, Osterloh and Frey provide a number of anecdotal examples of the use of JIF in contexts where it is inappropriate, and suggest that its use is widespread and increasing. Based on the evidence presented, it is equally possible that we have passed 'peak Journal

Impact Factor' and that its use is falling. It would be valuable to have empirical evidence about which of these perceptions is more accurate.

Moving to their second suggestion: if JIF is such a terrible heuristic, why is it still used? Osterloh and Frey suggest its continued use is because it makes people look better – the fact that citation distributions are heavily skewed means the JIF value is much higher than the actual number of citations earned by most articles. This is certainly true, but in any comparative context, most other people's JIF scores will be bigger, too. Hence, any comparative advantage is lost, and whose JIF score is bigger becomes something of a lottery. Not a complete lottery, since JIF does contain some information about number of citations, but unfortunately the authors do not quantify how much of a lottery, so we do not know how often the conclusion will be wrong for a particular ratio of JIFs. Osterloh and Frey show the overlap of citation distributions in Figure 1 of their paper – in this example it is clear that comparing the JIF values of the journals would generally tell you correctly which paper had the higher citation level, although there are exceptions – i.e., the heuristic is generally right but sometimes wrong.

It is also worth noting that, in the period immediately after publication, article level metrics are strongly stochastic, so JIF may be a better predictor of eventual citation level, and hence a better heuristic for quality, than article level metrics. An expert reading the publications could probably make a better assessment still, but such an expert may not be available or may not have the necessary time to do this properly.

It is equally plausible that JIF is still used because it is quick and because many researchers and administrators do not have easy access to article-level metrics, while failing to understand the pitfalls of JIF. JIF is easily and freely available, as well as being relatively stable, so you only have to look it up once for each journal. Obtaining normalised citation data has previously required access to proprietary databases and still requires looking up each and every paper. The non-heuristic option of reading every paper is enormously more time-consuming and requires considerable expertise to make a robust assessment.

My suggestion that we have passed ‘peak JIF’ is based on my 15 years of experience working with many researchers, administrator and funders across the world, so it is a little stronger than anecdote, while my suggestion that ease of use drives JIF usage is just a hypothesis. Others have suggested hypotheses drawing more on economic theories of scarcity and the tragedy of the commons (Casadevall & Fang, 2014). Unfortunately, Osterloh and Frey provide no evidence suitable for distinguishing which of these suggestions might be more accurate.

Putting these disagreements on one side, Osterloh, Frey and I agree that JIF is nearly always a bad heuristic. How might we prevent its use? There are two ways to stop something, namely carrots and sticks – in this case, either providing a better alternative or making JIF completely unpalatable.

Osterloh and Frey suggest a wholesale overhaul of the academic publishing model in order to weaken the link between JIF and quality (and hence article-level metrics) by incorporating a lottery element into whether borderline papers are accepted for publication in a particular journal. However, to significantly affect the relationship between JIF and quality, such a change would have to be widely adopted across academic publishing – and the authors suggest no incentive for a journal to incorporate a lottery element if competing journals are not doing so. Therefore, as a method of reducing JIF usage, this idea seems to be a complete non-starter. Given our agreement that JIF is already a terrible proxy for quality and yet people still use it, why should making the proxy a little worse change behaviour? Indeed there is good evidence that the strength of the relationship between JIF and median citation levels in journals is already falling and has been since 1990 (Lozano et al., 2012). None of this is to say that randomisation is not a good idea: Oswald (2020) makes a strong argument that incorporating randomisation in the selection of papers for publication would be of benefit to science, and others have suggested using or experimenting with its use in funding allocation (Avin, 2018; Barnett, 2016; Brezis, 2007; Fang & Casadevall, 2016; Guthrie, Ghiga & Wooding, 2017; Roumbanis, 2019); but it is hard to see how it could be the decisive factor in reducing the use of JIF (Oswald, 2000).

Heuristics, well used, are about focussing your attention where it can be most valuable. In a recruitment example – removing those candidates that have very little chance of being selected, and identifying those most likely to be recruited, means that you can then focus your detailed examination on the ones at the borderline. Long-listing, short-listing, first interview, second interview – each stage aims to focus more attention where it will most effectively improve decision-making. A danger of destroying commonly used heuristics is that people will fall back on using even worse ones. Using JIF to arrive at a 'long list' from a broadly distributed call for applications may be better than restricting the distribution of the call for applications through your networks, or long-listing from only 'good institutions' or researchers who have published, say, in *Science* and *Nature*.

There is general agreement that JIF is an almost universally bad heuristic and we should work to reduce its use. However, we need empirical evidence on how bad and how it compares to other possible approaches. Osterloh and Frey (2020) suggest the novel reason of 'borrowed plumes' (the fact that JIF is generally higher than raw citation numbers for each journal article) to explain why JIF is still used, and they come up with a radical suggestion of completely overhauling academic publication as a way to reduce its use. My assessment is that ease of use and lack of understanding are more important reasons for the continuing popularity of JIF. This would suggest a more feasible approaches to decreasing its use: firstly, to support research and initiatives that highlight the problems of JIF (Lozano et al., 2012) and build support for reducing its use through initiatives like the San Francisco Declaration on Research Evaluation (American Society for Cell Biology, 2012) and the Leiden Manifesto (Hicks et al., 2015); secondly to improve and encourage the use of normalised article-based citation metrics (Hutchins et al., 2016; Various, 2016; Waltman et al., 2011); and thirdly to support attempts to make such metrics more easily available (Bode et al., 2018; Thelwall, 2018).

If you want people to stop using a heuristic, you need to ask what they are using it for, why they are using it, and to understand what their other options are. We agree that JIF use needs to be curbed. Our difference of opinions about trends in JIF use and the best way to reduce it should be settled by good evidence on whether its use is increasing or falling;

where and why JIF is still used; and by testing and evaluating different approaches to curb JIF use.

Declaration of Competing Interest

I have worked for various philanthropic foundations and government agencies providing advice on research evaluation and research policy over many years.

Acknowledgements

With thanks to the Editor of this (virtual) Special Section, Ben Martin, for useful pointers and suggestions.

References

American Society for Cell Biology. (2012). San Francisco Declaration on Research Assessment (DORA). Retrieved August 3, 2019, from https://sfdora. org/read

Avin, S. (2018). Mavericks and lotteries. *Studies in History and Philosophy of Science Part A* (forthcoming). Published online at http://doi.org/10.1016/j.shpsa.2018.11.006

Barnett, A. G. (2016). Funding by lottery: political problems and research opportunities. *mBio*, *7*(4), e01369–16. http://doi.org/10.1128/mBio.01369-16

Bode, C., Herzog, C., Hook, D. & McGrath, R. (2018). *A Guide to the Dimensions Data Approach* (pp. 1–26). Digital Science. Retrieved from https://www.digital-science.com/resources/portfolio-reports/a-guide-to-the-dimensions-data-approach/

Brezis, E. S. (2007). Focal randomisation: An optimal mechanism for the evaluation of R&D projects. *Science and Public Policy*, *34*(10), 691–698. http://doi.org/10.3152/030234207X265394

Casadevall, A., & Fang, F. C. (2014). Causes for the persistence of impact factor mania. *mBio*, *5*(2), e00064–14–e00064–14. http://doi.org/10.1128/mBio.00064-14

Fang, F. C. & Casadevall, A. (2016). Research funding: the case for a modified lottery. *mBio*, *7*(2), 434. http://doi.org/10.1128/mBio.00422-16

Garfield, E. (2006). The history and meaning of the Journal Impact Factor. *JAMA*, *295*(1), 90–93. http://doi.org/10.1001/jama.295.1.90

Guthrie, S., Ghiga, I., & Wooding, S. (2017). What do we know about grant peer review in the health sciences? *F1000Research*, *6*, 1335. http://doi.org/10.12688/f1000research.11917.1

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S. & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature News*, *520*(7548), 429–431. http://doi.org/10.1038/520429a

Hutchins, B. I., Yuan, X., Anderson, J. M. & Santangelo, G. M. (2016). Relative Citation Ratio (RCR): a new metric that uses citation rates to measure influence at the article level. *PLoS Biology*, *14*(9), e1002541. http://doi.org/10.1371/journal.pbio.1002541

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323. http://doi.org/10.2307/24529203

Lozano, G. A., Larivière, V. & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for*

Information Science and Technology, *63*(11), 2140–2145. http://doi.org/10.1002/asi.22731

Martin, B. R. (2016). Editors' JIF-boosting stratagems – Which are appropriate and which not? *Research Policy*, *45*(1), 1–7. http://doi.org/10.1016/j.respol.2015.09.001

Osterloh, M. & Frey, B. (2020). How to avoid borrowed plumes in academia. *Research Policy*, 49, 103831 (this virtual Special Section).

Oswald, A. J. (2020). Rational randomization by journal editors – a mathematical derivation: a response to Osterloh and Frey's Discussion Paper on 'Borrowed Plumes'. *Research Policy* (this virtual Special Section).

Roumbanis, L. (2019). Peer review or lottery? A critical analysis of two different forms of decision-making mechanisms for allocation of research grants. *Science, Technology and Human Values*, *2*(1), 016224391882274. http://doi.org/10.1177/0162243918822744

Thelwall, M. (2018). Dimensions: A competitor to Scopus and the Web of Science? *Journal of Informetrics*, *12*(2), 430–435.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*(4157), 1124–1131. http://doi.org/10.1126/science.185.4157.1124

Various (2016). Special Section on Size-independent indicators in citation analysis. *Journal of Informetrics*, *10*(2), 329–684.

Wallace, M. L., Larivière, V. & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics*, *3*(4), 296–303. http://doi.org/10.1016/j.joi.2009.03.010

Waltman, L., van Eck, N. J., Van Leeuwen, T. N., Visser, M. S. & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47. http://doi.org/10.1016/j.joi.2010.08.001

Wouters, P., Sugimoto, C. R., Larivière, V., McVeigh, M. E., Pulverer, B., de Rijcke, S., & Waltman, L. (2019). Rethinking impact factors: better ways to judge a journal. *Nature*, *569*(7758), 621–623. http://doi.org/10.1038/d41586-019-01643-3

Yaqub, O. (2020). JIFs, giraffes, and a diffusion of culpability: a response to Osterloh and Frey's Discussion Paper on 'Borrowed Plumes'. *Research Policy*, (This Issue).