A simple proposal for the publication of journal citation distributions

Vincent Lariviere, Veronique Kiermer, Catriona J MacCallum, Marcia McNutt, Mark Patterson, Bernd Pulverer, Sowmya Swaminathan, Stuart Taylor, Stephen Curry

bioRxiv 062109; doi: http://dx.doi.org/10.1101/062109

Response to commentary and criticism of version 1

We are extremely gratified that the first version of this preprint, posted on 5th July 2016, has been so widely read (the PDF has been downloaded more than 11,000 times) and that it has generated such extensive commentary in a variety of outlets. These include comments on bioRxiv, on Pubpeer.com, in various blog posts (summarised here by Altmetric) as well as specific comments made on Twitter or received by email.

We thank all those who have taken the time and effort to provide critical feedback, which has been used to inform our revision of the preprint.

Many of the comments received made overlapping points. Rather than address every single one, we highlight here the most substantive criticisms and provide our response, indicating where appropriate, how they have been addressed in the revised preprint. The comments are grouped thematically. Responses to comments (where they appear) are indented.

For ease of navigation of this document, we have colour-coded the text. Comments are in **dark red**, while our responses are in **black** typeface.

Statistical analyses and methodology

JOHN SACK http://biorxiv.org/content/early/2016/07/05/062109 - comment-2787361297

Did you consider the (seemingly) simpler approach of calculating the skew (mean/median to keep it simpler) and having that number appear alongside any JIF? It would seem that a high skew would indicate the sort of 'lopsided' distribution that is worth noting. Advantages of this: it is easy to calculate, can be reported as a number (rather than a picture), and (because it is a number) can be computed and listed in tables such as your table 1. I have used the skew for this purpose myself.

An alternative would be to report the percentage of articles below the mean/JIF, which shows up several times in the text of the paper, suggesting that it communicates well exactly the phenomenon that is of concern.

> RICHARD SEVER http://biorxiv.org/content/early/2016/07/05/062109 - comment-2791541292
>
> I had the same thought. Since bioRxiv allows revised versions of manuscripts to be posted, perhaps the authors could revise the manuscript to include medians and interquartile ranges, along with a call for these to be promoted too. They might also consider adding pre-normalized versions of the distributions presented with the same y-axis scale as well.

ALEKEY BELIKOV http://biorxiv.org/content/early/2016/07/05/062109

I would highly suggest you to mention individual research metrics such as the h-index and others. It is completely ignored in your article, as if there are no alternatives to JIFs.

ALEKSEY BELIKOV http://biorxiv.org/content/early/2016/07/05/062109 - comment-2777594595

This reminds me of my recent manuscript on cancer incidence approximation by various probability distributions: http://www.biorxiv.org/content/early/2016/06/27/060970. You may try the same procedure to find which distribution best describes citations in a journal. It looks to me like log-normal, log-logistic or gamma distribution. Then each distribution can be described by only two or three parameters. Now THAT may be the replacement for JIF.

Response: Several other commentators (including additional remarks from Aleksey Belikov, and Phil Davis at the Scholarly Kitchen Blog) also suggested that alternative or additional metrics or parameterisation of the citation distributions would be useful substitutes for publication of the citation distributions themselves.

There are a number of points to make in response. First, we would like to emphasise that our proposal is not to replace the JIF with citation distributions, but to ensure that this information is published *alongside* the JIF wherever this is presented by journal publishers, in order to draw attention to the variation and spread in the data underlying the JIF. That being the case, use of other aggregate metrics is likely to fall into the same trap as the JIF, namely that they may conceal the full extent of features visible in the citation distribution. Second, we do not in principle object to the use of additional parameters to characterise the distributions but do not want to be prescriptive about what those should be. Even if journals opt to present this information, we would still recommend that the full distribution also be shown.

Phil Davis also suggested that our use of variable vertical scales could be problematic, given the wide variation in publishing volumes of different titles. We don't see this as particularly troublesome as long as the variation in vertical scales is clearly indicated (as in our Fig. 1). But if there is a desire to make comparisons between distributions, one way to address this would be to follow example given in Fig. 4b, where citation counts have been recalculated as percentages. Alternatively, the suggestion was made on Twitter by Rui Ponte Costa to generate kernel density estimates, which replot the citation data as estimates of the probability of citation. However, this cannot be done within Excel (an add-on is needed) and therefore introduces a level of complication that may inhibit uptake of our proposal.

To address these points, we have added a sub-section on Data Presentation to the Methods and a further comment on parameterization to the Discussion (3rd paragraph, beginning "Arguably, an alternative approach would be for journals…").

HAMED SEYED-ALLAEI http://biorxiv.org/content/early/2016/07/05/062109 - comment-2772676843

I have a suggestion regarding Fig 4. This figure is the spotlight of your work. But it is noisy, especially at the tails. This is natural, because there are few highly cited works. This can be improved using one of the following methods:

1. You can use logarithmic bins to construct the histograms: 0,1,2,4,8, ...

2. You can use cumulative density/histogram instead.

This reduces noises at the tails of the distributions so one can compares the performance of journals around highly cited works.

> DAVID COLQUHOUN http://biorxiv.org/content/early/2016/07/05/062109 - comment-2773718593
>
> Not so sure about this, for two reasons. Using logarithmic bins is not the same thing as looking at the distribution of log(citations) - something that's well known in the single ion channel world. And using cumulative distributions does not "reduce the noise", it merely conceals it. It's rarely a good idea.

Response: In line with our response to the points above, we feel that publication of the full distribution, even if it is noisy, is preferable since the noise itself signals the stochastic nature of citation patterns. It has also been pointed out (again by Phil Davis) that variations in binning of the citation counts will affect the appearance of the distributions. We would recommend that journals publish distributions without excessive binning of the data, and ideally, as we have done in our Fig. 1, using a binning interval of 1, to provide maximum resolution.

In the revised preprint these point have been addressed as described in our response to the preceding comment.

Technical questions and requests for clarification

CHRISTINA K PIKAS http://biorxiv.org/content/early/2016/07/05/062109 - comment-2770066110

I'm not clear on "key" - does this refer to the accession number for the article or is it something else found only in the paid version? Also, in appendix 1, why not do an index browse to find the journal? if you're looking for variations, also need to truncate?

> STUART TAYLOR http://biorxiv.org/content/early/2016/07/05/062109#comment-2772366484
>
> Thanks for the suggestion, Christina. We are looking into whether using the index feature in WoS/Scopus will generate a more reliable hitlist of articles than searching using title or ISSN. Whichever search mode is used, we recommend cross-checking with the journal's own published record.

Response: As described in the methods section, the 'key' is matching key used by Thomson Reuters in their Web of Science database to define links between citing and cited papers.

STEVE ROYLE - https://twitter.com/clathrin/status/750572427822960642

Fig 3 would be better with 1 pt lines, rather than markers (which obscure the other distributions).

Response: We agree and have amended Fig. 3 accordingly.

Points raised about the Discussion and Conclusions

ADAM EYRE-WALKER http://biorxiv.org/content/early/2016/07/05/062109 - comment-2784461666

I find there is a strange disconnect in arguments about the IF. The journal IF must contain some information about the merit of the papers published in a journal because we, the scientific community, are the ones that determine where things get published and what gets cited. We don't publish any old paper in *Nature* and *Science*; we publish what we believe is the best and most interesting science. Now sometimes, maybe even often, we will get this wrong, but an informed decision is made to publish a paper in a particular journal. In a sense all the IF represents is someone else's opinion about the merit of a paper. I think this might be one of the reasons people are uncomfortable with the IF along with the fact that the IF is clearly subject to error as a measure of merit. However, all measures of merit are subject to error and there is no evidence that the IF is any worse (http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001675). I'm not suggesting that the IF should be used blindly to assess papers and researchers, but suggesting that it contains little or no information about the merit of a paper seems illogical to me.

Response: This is a thorny issue but ultimately we disagree. The JIF refers to the *average* citation performance of papers in a given journal and has repeatedly been shown to be a poor predictor of the citation performance of individual pieces or work [e.g. Seglen, P.O. From bad to worse: evaluation by Journal Impact. *TIBS*, **14**, 326-7 (1989)].

But clearly this argument has struggled to convince. Eyre-Walker's comment points to a recurrent theme in research assessment: the complex linkage between journal prestige and the various reasons behind authors' choices of publishing venues for their own work and their decisions to cite the work of others. Our aim in proposing the publication of citation distributions is not to assert that citations counts are a better measure. In this regard, the use of *Nature* and *Science* in the comment above is in fact an example of the over-simplification that use of journal names or brands readily brings to processes of evaluation since these, being among the most prestigious journals in the world, are outliers that are not representative of the larger body of research literature. More commonly, judgments are being made between papers in journals where differences in JIF or reputation may not be significant (as pointed out in a recent commentary by Jeremy Berg).

The name of the journal where a paper is published and the numbers of citations that it attracts might both reasonably be thought of as interesting pieces of information in assessing a piece of work, but we would argue strongly that the process of assessment has to go beyond mere branding and numbers.

To address these points, in the Discussion section of the revised preprint, we have amplified our comments on the uses of the JIF and the difficulty in relating it meaningfully to assessments of individual pieces of work (Paragraph 3 beginning "We think that the variation evident…" and paragraph 6 beginning "Despite the overlap…").

LUDO WALTMAN http://biorxiv.org/content/early/2016/07/05/062109 - comment-2777455382

In addition to the comments made in the blog post, I also would like to raise the following issue.

In my view, the skewness of citation distributions can be interpreted in different ways, with different implications for the use of impact factors. Let me give two interpretations:

(1) This interpretation starts from the idea that citations provide a reasonable reflection of the quality of papers. Therefore the fact that within a single journal there are large differences in the

number of citations received by papers indicates that there are large differences in the quality of papers. Consequently, the impact factor of a journal doesn't properly reflect the quality of individual papers in the journal.

(2) This interpretation combines two ideas. The first idea is that citations are weak indicators of the quality of papers. Papers of similar quality on average have a similar number of citations, but there is a large standard deviation. Due to all kinds of 'distorting factors', papers of similar quality may differ a lot in the number of citations they receive. The second idea is that journals manage reasonably well to carry out quality control. Therefore the papers published in a journal are of more or less similar quality, so the standard deviation of the quality of the papers in a journal is relatively small. It follows from these two ideas that the impact factor, which is the average number of citations of the papers in a journal, provides a reasonable reflection the quality of individual papers in the journal (especially if the journal is sufficiently large, so that the above-mentioned 'distorting factors' in the citations received by individual papers cancel out). The fact that some papers in a journal receive many more citations than others is not the result of quality differences but instead it results from citations being weak indicators of quality, so it results from the above-mentioned 'distorting factors'. In this interpretation, impact factors are a stronger rather than a weaker indicator of the quality of individual papers than citation counts.

The interpretation that the authors seem to follow in their paper, and that for instance also seems to be followed in the DORA declaration, is the first one. However, the empirical results presented by the authors, showing that citation distributions are highly skewed, are compatible with both interpretations provided above. In the second interpretation, there is no reason to reject the use of IFs to assess individual papers in a journal. Therefore, if the authors want to reject the use of IFs for this purpose, I believe they need to provide an additional argument to make clear why the first interpretation is more reasonable than the second one. I do think that the first interpretation is indeed more reasonable than the second one, but a careful argument is needed to make clear why this is the case and on which assumptions this is based.

My comments are about the arguments that you use to support your ideas. In your paper, you for instance write: "Our intention here is to encourage publishers, journal editors and academics to generate and publish journal citation distributions as a countermeasure to the tendency to rely unduly and inappropriately on JIFs in the assessment of research and researchers."

You also write: "The distributions reveal that for all journals, a substantial majority of papers have many fewer citations than indicated by the arithmetic mean calculation used to generate the JIF and that for many journals the spread of citations per paper varies by more than two orders of magnitude. Although JIFs do vary from journal to journal, the most important observation as far as research assessment is concerned, and one brought to the fore by this type of analysis, is that there is extensive overlap in the distributions for different journals. Thus for all journals there are large numbers of papers with few citations and relatively few papers with many citations. This underscores the need to examine each paper on its own merits and serves as a caution against over-simplistic interpretations of the JIF."

How should these two quotes be understood? On the one hand, you put a lot of emphasis on the skewness of journal citation distributions, and on the other hand, you mention "the tendency to rely unduly and inappropriately on JIFs in the assessment of research and researchers". Based on this, my interpretation of your paper is that for you the inappropriateness of the use of IFs in research

evaluations follows, at least partly, from the skewness of journal citation distributions. This is how I read your paper, but please correct me if this is not how it is intended to be read.

At the same time, you also write: "citation counts cannot be considered as reliable proxies of the quality of an individual piece of research". If citations do not relate to quality, why then is it a problem that citation distributions are skewed?

Just to be clear about my position: It is not my intention to defend the use of IFs in research evaluations. My position is that, if one criticizes the use of IFs in research evaluations, the argument that one uses should be explained very carefully and should be fully clear and consistent. In my view, your argument doesn't yet satisfy these criteria.

Response: Waltman's comments raise several very interesting points and provide a useful theoretical framework for considering the properties and limitations of JIFs. However, they do not appear to resolve problem of *how* they should be used in research evaluation. He provides two suggested interpretations of the skew of citation distributions – that it arises (i) because citations reflect quality and quality is variable within any one journal, or (ii) from 'distorting factors' that mask the ability of journals to select and publish papers of 'similar quality'. Both interpretations contain elements of truth but both are also idealisations that may be difficult to dissect in reality. For example, it seems more plausible to suggest that journals select papers that they consider to meet a *minimum threshold* of quality or significance (in the judgement of editors and/or reviewers). Moreover, it is hardly controversial to suggest that while citations may well contain signals about quality of significance, they cannot be used as wholly reliable proxies for these properties. The correct interpretation of the skew of citation distributions therefore lies at some intermediate (and indeterminate) point, since the balance between these factors cannot easily be quantified and is likely to vary for individual papers in the same journal, and also for papers of similar quality in different journals.

Our argument is that over-reliance on the JIF in research assessment obscures this complexity and that publication of citation distributions is an aid to redirect attention. It is not a substitute for the JIF or an alternative metric – simply a reminder that consideration of JIFs or citations in relation to a particular piece of work is merely a starting point for further investigation.

In the Discussion section of the revised preprint, we have attempted to address these issues by adding further remarks on the complexities of the distributions and of their interpretation. (Paragraph 3 beginning "Arguably, an alternative approach would be for journals…" and paragraph 6 beginning "Despite the overlap…").

LUDO WALTMAN – https://www.cwts.nl/blog?article=n-q2w2c4 (Blog post – "The importance of taking a clear position in the impact factor debate")

Larivière et al. argue that "research assessment needs to focus on papers rather than journals" and that the IF "is an inappropriate indicator for the evaluation of research or researchers". On the other hand, Larivière et al. also state that they "are not arguing that the journal IF has no value in the comparison of journals". Hence, according to Larivière et al., IFs can be used for making comparisons between journals, but not for making comparisons between individual papers and their authors.

Larivière et al. reject the use of IFs to compare papers and their authors, and therefore I expect them to also reject the use of IFs in the above two examples. However, when Larivière et al. state that they "are not arguing that the journal IF has no value in the comparison of journals", what kind of use of IFs to compare journals do they have in mind? There is a strong interrelatedness of the use of IFs at the level of journals and at the level of individual papers. A simple statement that IFs can be used at the former level but not at the latter one therefore doesn't seem satisfactory to me.

In discussions on IFs, there are two clear positions that one could defend. On the one hand, one could take the position that in certain cases the use of IFs at the level of journals and at the level of individual papers is acceptable. On the other hand, one could take the position that any use of IFs should be rejected. Although these positions are opposite to each other, they each seem to be internally consistent. Larivière et al. take neither of these positions. They argue that the use of IFs at the level of individual papers should be rejected, while the use of IFs at the level of journals is acceptable. This seems an ambiguous compromise. Given the strong interrelatedness of the use of IFs at the two levels, I doubt the consistency of rejecting the use of IFs at one level and accepting it at the other level.

To have a fruitful debate on the IF, it is essential that everyone involved takes a position that is clear and internally consistent. Critics of the IF, such as Larivière et al. but also for instance the supporters of the San Francisco Declaration on Research Assessment, need to accept the full consequences of their criticism. Rejecting the use of IFs at the level of individual papers seems to imply that there also is little room for the use of IFs at the level of journals. In order to be consistent, critics of the IF may even need to further extend their criticism. If one rejects the use of IFs because of the variability in the quality of the papers in a journal, this calls into question whether other types of information on the quality level of journals, such as researchers' personal experiences with journals, can still be used. Shouldn't the use of this information be rejected as well? For instance , when deciding which paper to read or which colleague to collaborate with, shouldn't one completely ignore any information, both from personal experience and from quantitative indicators, on the quality level of the journals in which papers have appeared? This may seem quite an extreme idea, but at least it represents a clear and consistent position, and in fact some have already taken concrete steps to move in this direction.

Response: In our view this critique over-states our position. We have not argued that "the use of IFs at the level of individual papers should be rejected", rather that over-simplistic use of JIFs in assessing individual papers is not defensible and that publication of citation distributions helps to *draw attention* to the complexity of the data underlying the JIF and should serve as a prompt for deeper investigation of the merits of a particular piece of research. We also question the assertion of "the strong interrelatedness of the use of IFs at the two levels" (meaning at the level of the journal and the level of the paper), since this has long been questioned [*e.g.* Seglen, P. O. Why the impact factor of journals should not be used for evaluating research. *BMJ* **314,** 498–502 (1997)].

Nevertheless, in light of this comment we recognize the need to clarify some of the finer distinctions that we have attempted to draw in our preprint, in particular what we meant by stating that we "are not arguing that the journal IF has no value in the comparison of journals". These clarifications incorporated in the modifications to the Discussion mentioned above but particularly the paragraph beginning "Despite the overlap,…".

On the value of citations as indicators of quality

DAVID COLQUHOUN http://biorxiv.org/content/early/2016/07/05/062109.article-metrics-comment-2773741965

If one considers actual individual publications, it soon becomes clear that citations are a useless way to assess quality Just look at the huge number of citations to Andrew Wakefield's (falsified) paper in the Lancet.

The same is true of my own publications. For example, a recent unoriginal and trivial review has had 125,000 full text views, 18,000 pdf downloads and 70 citations in a year and a half -see http://rsos.royalsocietypublishing.org/content/1/3/140216.

In contrast, really original (but rather mathematical) work, like http://www.onemol.org.uk/Colquhoun-Hawkes-Srodzinski%201996-ocr.pdf has had barely more citations in 20 years. The size of the potential audience, and lack of hard maths, are the most important things in determining citations. Quality has very little to do with it.

> ALEKSEY BELIKOV http://biorxiv.org/content/early/2016/07/05/062109.article-metrics-comment-2775112570
>
> Nice to hear from you, I have seen that p-value paper before, it is the most read paper in that journal. Congratulations! I think you could have tried Nature or Science with that paper, if you haven't. Although I agree with you that citations are not the direct indication of article quality, and articles are best judged by reading them, I insist that for cases when reading is not possible (e.g., when you have to sift through 1000s of papers of 100s of applicants), citation-based metrics are the best option available. Allow me to reproduce my comment to Science editorial (http://comments.sciencemag.org/content/10.1126/science.aaa3796#comments) in defense of citation-based metrics:
>
> "The use of citation metrics for evaluating researches has been criticized heavily by many, including this article by the Editor-in-Chief of Science. The main argument is that citations do not capture the essence of research excellence. Let me try to defeat this argument. To understand what a citation means, we need to reflect on our own behavior as scientists when actually citing something. First time we use citations is in the introduction to a paper. Everyone knows that a good practice is to mention all publications directly relevant to the article. They do not necessarily need to have had a strong influence on our research though, but they are useful for readers to understand the findings in a context. Citations in the discussion section are typically used to compare findings of the paper to that in the field, thus performing the same context-setting function. Finally, citations in the results and methods parts are rare and usually refer to protocols and techniques. Thus, the majority of citations link the article with similar ones in the field. Why then some articles receive orders of magnitude more citations than others? Employing the context-setting concept introduced above, it can be said that those articles contribute to the context of many more articles than an average article, i.e. they are central to the field. The most-heavy cited articles transcend the boundaries of a specific field and create the context for the whole of biomedicine, physics, or even science. Thus, it is not by chance that top journals state "influential across fields", "interesting to an interdisciplinary readerships" or "merit recognition beyond that

provided by specialty journals" as their selection criteria. To discover something so broadly important is not easy and thus requires research excellence. Closing the circle, citations, citation counts, and citation-based indices have not fallen from the sky. And they are here to stay."

P.S. Your p-value paper is heavily cited because it raises the issue that affects almost all of the science.

ALEKSEY BELIKOV http://biorxiv.org/content/early/2016/07/05/062109.article-metrics-comment-2775097850

Well, I think both single numbers and distributions are needed. Distributions cannot be easily compared - you will anyway need to calculate the median or mode from it, which are single numbers. And it is not clear whether they really describe these distributions in a best possible way. When you look at the long lists of journals, or scientists, and want to sort them by impact, how would you do this without a single number for each journal/scientist? As you showed in your paper, current technology easily permits to create distributions to anyone interested in particular journal or scientist, but for quick sifting through large lists of candidates single numbers are unbeatable.

ANON https://pubpeer.com/publications/38118904A819344EF9D758C8A431A8

More seriously, the method assumes that the number of citations is correlated with the quality and scientific usefulness of the work. I think any such correlation must also be quite weak. One can certainly find examples of highly cited work later discovered to have been fabricated - nobody can have built on that.

Response: We agree with Colquhoun that citations cannot be regarded as a straight-forward indicator of the quality of the cited work because the reasons for citation vary (as already discussed above). In many cases of course citation arises because of the interest or significance of a piece of work, but in other cases it may be to discount or refute findings. We would argue that citations provide an interesting signal that requires further investigation before their meaning can be validated.

We have added a comment to this effect in the Discussion paragraph beginning "We think that the variation evident in the distributions…"

RICHARD MALHAM http://biorxiv.org/content/early/2016/07/05/062109.article-metrics-comment-2771880067

I already noted these in a comment on this blog post (http://blogs.plos.org/plos/2016/07/impact-factors-do-not-reflect-citation-rates/) but wanted to also raise them here

Stephen: When revising the article, I would suggest looking at the existing recommendations in these two reports (particularly relevant to the recommendations section at the end of the Discussion section)

'Evaluating Interdisciplinary Research: a practical guide', from Durham University's Institute of Advanced Study (July 2015). https://www.dur.ac.uk/ias/news/?itemno=25309

'Improving recognition of team science contributions in biomedical research careers', from the Academy of Medical Sciences (March 2016). http://www.acmedsci.ac.uk/policy/policy-projects/team-science/

Response: We thank the commenter for drawing our attention to these studies. In the revised preprint we have amplified our discussion of work elsewhere that has sought to develop more holistic to research assessment, particularly at disciplinary interfaces.

We have added mention of this work in the new "Conclusions and Recommendations" section that follows the Discussion, in the 1st paragraph, beginning "The co-option of JIFs…".

SARAH DE RIJCKE https://www.cwts.nl/blog?article=n-q2x234 (blog post – "Let's move beyond too simplistic notions of 'misuse' and 'unintended effects' in debates on the JIF")

[…] Though I applaud sincere, methodologically sophisticated calls for more transparency such as the one made by Larivière et al., I am afraid they do not suffice. The recourse we then take is towards an upstream solution, guided by an optimistic yet also slightly technocratic mode of 'implementation' (De Rijcke & Rushforth, 2015). If journals would indeed start to publish the citation distributions behind their JIFs, what exactly would this change on the shop-floor, in assessment situations, and in the daily work of doing research?

Larivière et al. put forth a methodologically driven plea to focus not on the JIF but on individual papers and their actual citation impact. Though commendable, I think this strategy obscures a much more fundamental issue about effects of the JIF on the daily work of researchers and evaluators. JIF-considerations have a tendency to either move to the background other measures of scientific quality (e.g. originality, long-term scientific progress, societal relevance), or to allow them to become redefined through their relations to the JIF and other quantitative performance indicators. In my opinion this insight leads to a crucial shift in perspective. For truly successful interventions into indicator-based assessment practices to happen, I think we need to move beyond too simplistic entry points to the debate of 'misuse' and 'unintended effects'. My hypothesis is that researchers (and evaluators) continue to use the JIF in assessment contexts - despite the technical shortcomings – for the complicated reason that the indicator is already so engrained in different knowledge producing activities in different fields. Our research findings suggest that in calling for researchers and evaluators to 'drop' the JIF, people are actually calling for quite fundamental transformations in how scientific knowledge is currently manufactured in certain fields. This transformation is the primary, and also the quite daunting, task.

We agree wholeheartedly with Dr De Rijcke that the underlying problem is that the JIF "is already so engrained in different knowledge producing activities in different fields." While we would point out that our proposal is not to focus on the "citation impact" of individual papers and that the publication of citation distributions will not be sufficient to eradicate a deeply embedded culture, we would argue that it is nevertheless an important step in advancing the conversation.

It is true that we have laid some emphasis on the current misuse of JIFs in evaluation, but this is re-iterating points made elsewhere [including in De Rijcke & Rushforth (2015)]. Nevertheless, we take

on board the broader point that solutions cannot be formulaic, not least because researchers are both subjects and users of JIFs at the present time.

We have therefore amplified our discussion of the context in which our proposal is offered to emphasise the difficulty of enacting cultural change. See paragraph beginning "The co-option of JIFs…" in the "Conclusions and Recommendations" section.

ALEX RUSHFORTH https://www.cwts.nl/blog?article=n-q2x234

The argument of Larivière et al's paper which started this debate seems to rest on a kind of path dependency-account – the scientific system is locked-in around an inferior tool when there are better alternatives available. Although I would agree with this, I also think path dependency accounts are not the only way forward in these kinds of debates – for my money they can be a bit too restrictive, as by suggesting salvation lies in adopting a superior tool they close-off a lot of what is going on in the research system which we should pay closer attention to. What I think is more important is to understand the kinds of conditions under which something like the impact factor can come to have such a big effect on the way research is conducted and governed – in interviews Sarah and I conducted with scientists this revolved around issues like scarcity of resources, excessive quantities of scientific literature from which to choose what to read (and cite), fiercely competitive job markets, the fact researchers are incentivized to write and not to read etc. It is confronting these sorts of issues (which studying the uses of the impact factor points us toward) more than technical limitations of the impact factor per se which I think would help produce better systems of evaluation and ultimately better science. I fear that getting journals to publish distributions alongside JIFs will not loosen the grip of the impact factor – it's not that researchers we spoke to aren't aware of limitations in how the indicator is calculated – it's more they recognize it's the de facto standard against which their prospect for external grants or a job interview will depend. It's these latter kinds of issues which will need to change first if the impact factor is to go away.

Response: As noted above, we agree that formulaic or technical proposals are not a wholly adequate solution to the problems thrown up by the use of JIFs in research evaluation. We agree also that many researchers are aware of some of the limitations of such indicators. However, that has not prevented instances of over-simplistic use by other researchers or by research managers. We hope that wider publication of citation distributions will succeed in drawing greater attention to the problems associated with inappropriate use of metrics and help to focus deliberations more on the work itself.

We agree above that this proposal in itself falls well short of being a solution. But it is part of the groundwork that is necessary in order to push all those involved in research assessment to tackling the underlying problem.

We have added remarks to this effect in the paragraph beginning "The co-option of JIFs …" in the "Conclusions and Recommendations" section.