THE ROYAL SOCIETY PUBLISHING

ROYAL SOCIETY OPEN SCIENCE

The missing metric: quantifying contributions of reviewers

Maurício Cantor, Shane Gero

Article citation details
R. Soc. open sci. **2**: 140540.
http://dx.doi.org/10.1098/rsos.140540

Review timeline

Original submission:	28 September 2014
Revised submission:	23 December 2014
Final acceptance:	13 January 2015

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSOS-140341.R0 (Original submission)

Review form: Reviewer 1 (Jelte Wicherts)

Is the manuscript scientifically sound in its present form?
Yes

Are the interpretations and conclusions justified by the results?
Yes

Is the language acceptable?
Yes

Is it clear how to access all supporting data?
Yes. Data were there!

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Major revision is needed (please make suggestions in comments)

Comments to the Author(s)
Review of "The missing metric: Quantifying contributions of reviewers" by Jelte M. Wicherts

In this manuscript the authors present a novel metric to assess the contribution of peer reviewers over time. I wholeheartedly agree with their statement that currently most journals do insufficiently incentivize the writing of high quality reviews, and that this may affect the quality of the peer review system. Many journals acknowledge reviewers by listing them at the end of each year, but most of the rewarding of reviewers is informal, with little effort to standardize. For instance, a regular reviewer with subjectively assessed reviews may be promoted to become a member of the editorial board or may become (action/associate/chief) editor him/herself over time. Such a reward system may benefit from having an index of peer reviewer's contributions that is (1) is intuitive, (2) comparable across journals, (3) fair to early-career scientists, and (4) contains ingredients about which most scientists (and/or editors?) agree. The current authors propose such an index, which they denoted R. I liked their idea in general and welcome debate about such an index. But for any index to become useful, it is vital that these four goals (and possibly others) are met. I focus my review on the debate whether R meets these requirements.

Let me compare, for the sake of discussion, R to the H-index that has so rapidly grown in popularity and use. One might criticize the H-index on several grounds. Specifically, H is incomparable across fields that differ in size, pace of publication, number of authors per article, and citation culture. Similarly, it is clearly unfair to early-career scientists and could be viewed as a measure of seniority and of a scientist's ability to acquire co-authorship. Yet, technically, H has some nice features. It is attempts to describe the distribution of highly skewed citation data. It is impossible to "game" by simply having a few highly-cited articles or simply by publishing a lot of papers that few are willing to cite. H has what psychometricians would call face validity, because it rings an intuitive bell with many users. The question is whether R has some of the same nice features as H does. I am not sure. The index R does not appear strong in terms of face validity and I was unable to arrive at an easy interpretation and its intellectual parents failed to provide it in the current manuscript. It has no clear "scaled value" that provides values that are readily interpretable (e.g. what does R of 50 mean?). In my view, this renders R less convincing to potential users, thereby lowering its chance of being adopted widely.

Surely, the former point does not take away other potential qualities of R, like its potential to contribute to the rewarding of high quality reviews. So we are left with points 2-4 above. The authors take into account journals' Impact Factor (IF) and want to use a measure of the quality of the individual reviews that is standardized, but it is not entirely clear how they want to achieve this. Also, it is not immediately clear how R deal with the fact that higher IFs are associated with lower word counts in some journals (Science and Nature have clear word limits, whereas some substantive journals with lower IFs have none). As the authors argue, R does appear to deal with point (3), which is probably a good thing given the research showing that early and mid-career scientists write the best reviews and are often (at least in my experience at PLOS ONE) more willing to review. Perhaps the most essential question is (4): would all (or at least most) stakeholders agree with the inclusion of w, s, and IF in the index? Here I have some doubts. I could envision a lot of my colleagues arguing vigorously against the adoption of the IF (although I am not as critical as they are regarding IF), and there may also be issues with w. For instance, I also review for technical journals in my field and these reviews typically require much more time (e.g., because I need to check the formulas) despite the fact that the number of words in the manuscripts is limited.

The core question about support for the inclusion of these ingredients is probably mostly an empirical one. Hence, I feel that the current work would be considerably strengthened by adding data on the support for the R index (and its ingredients) in various fields. Similarly, one would need to have more data in order to get a sense of the weighting that the current R index uses for IF, w, and s. The current rationale for the use of the current weights did not entirely convince me and needs more work.

Review form: Reviewer 2 (David Duffy)

Is the manuscript scientifically sound in its present form?
No

Are the interpretations and conclusions justified by the results?
No

Is the language acceptable?
No

Is it clear how to access all supporting data?
The data seem to be there but my Mac was quite hostile about opening the files. There does;t seem to be a metadata or index of the data, but then again it may be the Mac OS being snotty.

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
I do not feel qualified to assess the statistics

Recommendation?
Major revision is needed (please make suggestions in comments)

Comments to the Author(s)
The English needs work and the discussion needs to be tightened up.

I think the model can be better explained, with more details. More specifically, is the number of words in a paper really a proxy for time spent? And is this the published number? A reviewer may put a lot of time into helping shorten a paper but their zeal would in fact be penalized. Also some fields are wordier than others. For the s-score how does one standardize across journals (and across editors) and how does one weigh the various aspects of a review? Only whether the reviewer met the deadline is objective; the rest are not.

l. 76 early career reviewing "many". But until they are known and published, they will not be asked to be reviewers. This and what follows assumes a Peer of Science arrangement where reviewers know what manuscripts are out there and can chose to do lots of reviews. For many journals, reviewing is by invitation only.

l. 80-85 opportunist and specialist--what is the basis for this? Many might pursue a 'mixed' strategy. For example, a senior person might be both opportunistic, if offered the chance to review for Nature or Science, but willing to review lots of papers within their own field.

94 "tendency" within the model's conditions

98 how did you adjust for editor's feedback? This needs bait more explanation, as in l. 100 why would opportunistic reviewers provide poor reviews to top journals?

l. 110 "standing in the field" not clear how it does this

I guess I feel uneasy at the mixed results/discussion as it is hard to tell where results end and discussion begins, i.e. l. 111 "unheralded pillars", l. 113 :"in our experience", "we assume", l. 118 "hard working", l. 119 "typically", ;. 121 "daunting", l. 126 "inevitably" l. 144-145 results appear, after a long run of discussion. I'd suggest separating them and elaborating on the results and shortening the discussion.

l. 147. judging journals by metrics. Editors will game this, as they will score their own reviewers higher to elevate R (and to avoid annoying their reviewers who will avoid reviewing for journals with tough-scoring editors), unless there are quantitative standards.

l. 156 "practical aid in the publication process". One complaint of reviewers and editors is that sometimes authors expect them to do the work of honing the manuscript for publication. In other words, an author is seen as "just sending it in to see how the reviewers respond", rather than having colleagues vet the paper before submission.

Given the gaming that is going on with less reputable on-line journals, the authors should look into how such journals might be used to game R, as they are used to game H and other indices. Indeed they might tinker with the model and see what vulnerabilities it has.

I do wish the authors would have actually cited my paper rather than referring to it in the acknowledgements, but I am happy in any event that they are advancing the idea.

David Cameron Duffy, University of Hawaii Manoa.

Decision letter (RSOS-140341)

02-Dec-2014

Dear Mr Cantor:

Manuscript ID RSOS-140341 entitled "The missing metric: Quantifying contributions of reviewers" which you submitted to Royal Society Open Science, has been reviewed. The comments from reviewers are included at the bottom of this letter.

In view of the criticisms of the reviewers, the manuscript has been rejected in its current form. However, a new manuscript may be submitted which takes into consideration these comments.

Please note that resubmitting your manuscript does not guarantee eventual acceptance, and that your resubmission will be subject to peer review before a decision is made.

You will be unable to make your revisions on the originally submitted version of your manuscript. Instead, revise your manuscript and upload the files via your author centre.

Once you have revised your manuscript, go to https://mc.manuscriptcentral.com/rsos and login to your Author Center. Click on "Manuscripts with Decisions," and then click on "Create a

Resubmission" located next to the manuscript number. Then, follow the steps for resubmitting your manuscript.

Your resubmitted manuscript should be submitted by 01-Jun-2015. If you are unable to submit by this date please contact the Editorial Office.

I look forward to a resubmission.

Sincerely,
Emilie Aime
Senior Publishing Editor, Royal Society Open Science
openscience@royalsociety.org

Author's Response to Decision Letter for (RSOS-140341)

See Appendix A.

RSOS-140540 (Revision)

Review form: Reviewer 1 (Jelte Wicherts)

Is the manuscript scientifically sound in its present form?
Yes

Are the interpretations and conclusions justified by the results?
Yes

Is the language acceptable?
Yes

Is it clear how to access all supporting data?
Yes

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Accept as is

Comments to the Author(s)
The authors have revised their manuscript well and I have no further comments.

Review form: Reviewer 2 (David Duffy)

Is the manuscript scientifically sound in its present form?
Yes

Are the interpretations and conclusions justified by the results?
Yes

Is the language acceptable?
Yes

Is it clear how to access all supporting data?
Yes

Do you have any ethical concerns with this paper?
No

Have you any concerns about statistical analyses in this paper?
No

Recommendation?
Accept as is

Comments to the Author(s)
The authors have addressed the major issues. There are other items that inherently haven't as clear answers. We could tweak the formula but its effectiveness is best measured by how widely it is adopted rather than by continuing review. The authors have done the community a service with this effort.

Decision letter (RSOS-140540)

09-Jan-2015

Dear Mr Cantor

On behalf of the Editor, I am pleased to inform you that your Manuscript RSOS-140540 entitled "The missing metric: Quantifying contributions of reviewers" has been accepted for publication in Royal Society Open Science

The reviewers and Subject Editor have recommended publication, therefore please proofread your manuscript carefully and ensure that the following editorial points are addressed.

• Ethics statement
If your study uses humans or animals please include details of the ethical approval received, including the name of the committee that granted approval. For human studies please also detail whether informed consent was obtained. For field studies on animals please include details of all permissions, licences and/or approvals granted to carry out the fieldwork.

• Data accessibility
It is a condition of publication that all supporting data are made available either as supplementary information or preferably in a suitable permanent repository. The data accessibility section should state where the article's supporting data can be accessed. This section should also include details, where possible of where to access other relevant research materials such as statistical tools, protocols, software etc can be accessed. If the data has been deposited in an external repository this section should list the database, accession number and link to the DOI for all data from the article that has been made publicly available. Data sets that have been deposited in an external repository and have a DOI should also be appropriately cited in the manuscript and included in the reference list.

• Competing interests
Please declare any financial or non-financial competing interests, or state that you have no competing interests.
• Authors' contributions
All submissions, other than those with a single author, must include an Authors' Contributions section which individually lists the specific contribution of each author. The list of Authors should meet all of the following criteria; 1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; 2) drafting the article or revising it critically for important intellectual content; and 3) final approval of the version to be published.

All contributors who do not meet all of these criteria should be included in the acknowledgements.

We suggest the following format:
AB carried out the molecular lab work, participated in data analysis, carried out sequence alignments, participated in the design of the study and drafted the manuscript; CD carried out the statistical analyses; EF collected field data; GH conceived of the study, designed the study, coordinated the study and helped draft the manuscript. All authors gave final approval for publication.

• Acknowledgements
Please acknowledge anyone who contributed to the study but did not meet the authorship criteria.

• Funding statement
Please list the source of funding for each author.

Because the schedule for publication is very tight, it is a condition of publication that you submit the revised version of your manuscript within 7 days (i.e. by the 18-Jan-2015). If you do not think you will be able to meet this date please let me know immediately.

To revise your manuscript, log into https://mc.manuscriptcentral.com/rsos and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions". Under "Actions," click on "Create a Revision." You will be unable to make your revisions on the originally submitted version of the manuscript. Instead, revise your manuscript and upload a new version through your Author Centre.

When submitting your revised manuscript, you will be able to respond to the comments made by the referees and upload a file "Response to Referees" in "Section 6 - File Upload". You can use this to document any changes you make to the original manuscript. In order to expedite the processing of the revised manuscript, please be as specific as possible in your response to the referees.

When uploading your revised files please make sure that you have:

1) A text file of the manuscript (tex, txt, rtf, docx or doc), references, tables (including captions) and figure captions. Do not upload a PDF as your "Main Document".
2) A separate electronic file of each figure (EPS or print-quality PDF preferred (either format should be produced directly from original creation package), or original software format)
3) Included a 100 word media summary of your paper when requested at submission. Please ensure you have entered correct contact details (email, institution and telephone) in your user account

4) Included the raw data to support the claims made in your paper. You can either include your data as electronic supplementary material or upload to a repository and include the relevant doi within your manuscript
5) Included your supplementary files in a format you are happy with (no line numbers, vancouver referencing, track changes removed etc) as these files will NOT be edited in production

Once again, thank you for submitting your manuscript to Royal Society Open Science and I look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes
Emilie Aime
Senior Publishing Editor
openscience@royalsociety.org

Author's Response to Decision Letter for (RSOS-140540)

To the editorial board at the Royal Society Open Science Emilie Aime, Senior Publishing Editor Dear Dr. Aime, We are very happy with the editorial decision on the manuscript RSOS-140540. We thank the editor and the two reviewers for the final positive feedback on our work. Please find attached the final manuscript, its two figures in high-quality pdf format and the electronic supplementary materials (simulation details and the R package). We double-checked the manuscript to make sure it includes all the required sections: Author's contributions, Acknowledgements, Funding statement, Competing Interests and Data accessibility. We are making the original data and the simulated data available with the statistical package for R as an electronic supplementary material. The data is necessary to run the simulations in the package thus we found more straightforward to provide the whole material together. In addition, the same material is available in our online repository described in the "Data accessibility" section. Finally, the Ethics statement section does not apply to our manuscript. Once again, we thank you very much for managing our manuscript. Sincerely, Mauricio Cantor & Shane Gero

Appendix A

Manuscript ID RSOS-140341
The missing metric: Quantifying contributions of reviewers
Mauricio Cantor & Shane Gero

To the editorial board at Royal Society Open Science
December 22, 2014

Dear Dr. Emilie Aime,
Senior Publishing Editor

We are very grateful for the opportunity of reviewing and resubmitting our manuscript RSOS-140341. Both reviewers strongly supported the need for quantifying reviewers' contributions as a way to improve the peer-review system, while also raising valid points we needed to address. Please, see below a point-by-point response letter describing how we addressed each of their comments, both rephrasing and including new analyses, in the new version of the manuscript.

We believe strongly that any version of an index for such purpose is unlikely to completely satisfy the entire scientific community. This is made evident among these two reviewers, who welcome the index but disagree with each other about the parameters of the index. While they have questioned some of the proxies we have used to quantify reviewers' contributions, they have not proposed any alternatives. Moreover, the reviewers have not requested specific changes on the index, such as pointing out which particular parameters they would like to add or remove from the equation. Therefore we chose to keep the original formulation, but have greatly improved the justification for each parameter, both in the manuscript and in the response letter. To support the final formulation, we have reanalysed the original data and included additional analyses in the supplementary material (ESM3). We modelled 8 alternative, reduced versions of our index to show that removing parameters from the index might change its absolute scale, but it would not improve its validity, utility, or applicability. The additional parameters weight the number of reviewed manuscripts with other contributions of reviewers (time invested, standing in the field, quality of the review). This way, our proposal is more conservative and appropriate to capture the essentials of the contributions through peer-reviewing. Most critically from the standpoint of publication in *Royal Society Open Science*, our metric delivers on the first three key integrative components suggested by Dr. Wicherts. As for the last component, the debate upon if the metric will be accepted widely, it can only genuinely take place after it is formally proposed in press.

Encouraged by their positive feedback and enthusiasm for the need of evaluate and reward reviewers, we are resubmitting a new version of the manuscript RSOS-140341 for your consideration for publication in the *Royal Society Open Science.* This review process was not only a great opportunity to improve our work, but an proof that focus on measuring reviewers' efforts can indeed stimulate high-quality and thoughtful reviews. We hope that you will find our revised manuscript suitable for publication.

All our very best,

Mauricio Cantor
Department of Biology
Dalhousie University, Canada
&
Shane Gero
Department of Biosciences
Aarhus University, Denmark

Comments from the Reviewer 1, Dr. Jelte M. Wicherts

Dear Dr. Wicherts,
We very much appreciate the thoughtful signed review highlighting the strong points of our manuscript, as well as many important issues that begged more explanation. Please find below how each of your comments were addressed in the new version of the manuscript.

Reviewer 1:
In this manuscript the authors present a novel metric to assess the contribution of peer reviewers over time. I wholeheartedly agree with their statement that currently most journals do insufficiently incentivize the writing of high quality reviews, and that this may affect the quality of the peer review system. Many journals acknowledge reviewers by listing them at the end of each year, but most of the rewarding of reviewers is informal, with little effort to standardize. For instance, a regular reviewer with subjectively assessed reviews may be promoted to become a member of the editorial board or may become (action/associate/chief) editor him/herself over time. Such a reward system may benefit from having an index of peer reviewer's contributions that is (1) is intuitive, (2) comparable across journals, (3) fair to early-career scientists, and (4) contains ingredients about which most scientists (and/or editors?) agree. The current authors propose such an index, which they denoted R. I liked their idea in general and welcome debate about such an index. But for any index to become useful, it is vital that these four goals (and possibly others) are met. I focus my review on the debate whether R meets these requirements.

Response by the Authors: We thank the reviewer for clearly laying out the criteria for an ideal metric. Dr. Wicherts supports the need of such a metric and we feel that our metric can or has the ability to deliver on all four of the goals outlined. Furthermore, we have refocused part of our discussion to highlight how R-index fits into your outline of the ideal metric, by dedicating an entire subsection to it (starting on L229). We will highlight our specific comments below in relations to these goals.

Goal 1: Intuitive

Reviewer #1: Let me compare, for the sake of discussion, R to the H-index that has so rapidly grown in popularity and use. One might criticize the H-index on several grounds. Specifically, H is incomparable across fields that differ in size, pace of publication, number of authors per article, and citation culture. Similarly, it is clearly unfair to early-career scientists and could be viewed as a measure of seniority and of a scientist's ability to acquire co-authorship. Yet, technically, H has some nice features. It is attempts to describe the distribution of highly skewed citation data. It is impossible to "game" by simply having a few highly-cited articles or simply by publishing a lot of papers that few are willing to cite. H has what psychometricians would call face validity, because it rings an intuitive bell with many users. The question is whether R has some of the same nice features as H does. I am not sure. The index R does not appear strong in terms of face validity and I was unable to arrive at an easy interpretation and its intellectual parents failed to provide it in the current manuscript. It has no clear "scaled value" that provides values that are readily interpretable (e.g. what does R of 50 mean?). In my view, this renders R less convincing to potential users, thereby lowering its chance of being adopted widely.

Authors: We appreciate the insightful exercise of comparing the R-index with the well-stablished H-index. Our goal was to operationalize an index that is simple to calculate and yet contains multiple parameters to capture the reality of an individual's time and efforts as a reviewer. By doing so, the index gains in broader integrity but yields numerical outputs that could be immediately less intuitive—but not less informative—

if directly compared to the H-index. Put simply, the R-index quantifies reviewers proportionally to their contributions. It is a more relative comparator than the absolute face validity of H. The R-index increases almost linearly with the number of reviews, while it is weighted by other relevant aspects (time and effort invested; standing in the field) that prevent it from being gamed.

While an H-index of 50 means an author published 50 articles that were cited at least 50 times, an R-index of 50 means a solid contribution to the review system, which can be achieved through different routes. For instance, a R-index could result from a) 50 excellent reviews of long manuscripts to low rank journals (n=50, s=1, IF=1, w=10000); b) 100 excellent reviews of short manuscripts to low rank journals (n=100, s=1, IF=1, w=5000), c) 100 good reviews of short manuscripts to mid rank journals (n=100, s=0.5, IF=4, w=5000), d) 25 very good reviews of short manuscripts to top journals (n=25, s=0.8, IF=25, w=5000). The peer-review system is equally benefited from different individual contributions. Thus, as opposed to H-index, the R-index is more diversely applicable and egalitarian since it levels off different reviewers' styles, career stages and disciplines. As Dr. Wicherts point out, we failed to clearly interpret this in text and it now appears in the current revised manuscript (see the new section "*The ideal of an ideal metric*", L229. What is deemed by academia to be a "good" R-index is as controversial as what is deemed to be an impressive H. The publication of this proposed metric is partly meant to encourage this debate. In particular, as we highlight in the manuscript (L187-189), the ratio of H to R will be particularly revealing of a researchers contributions to the community. This ratio yields the number of constructive reviews produced for each high quality publication.

We would also like to highlight that, as Dr. Wicherts points out, the H-index suffers in several of the ideal criteria for a metric in so much as it delivers strongly on (1) but is weak on both (2) and (3). Although it is easy to interpret, it is not comparable across fields and is challenging to early career scientists. Nonetheless, even with its weaknesses, it has risen to virtual ubiquity and has spawned a number of variants which address directly some of its shortfalls. The H-index has created the substrate for discussion and change, arguably negative or positive, in our community in relation to researcher assessment. We believe that the publication of the index quantifying reviewers, and the discussion which follows, will do the same for our peer-review system. And it appears that on this point, Dr. Wicherts agrees with us.

Goal 2: Comparable across journals

Reviewer 1: Surely, the former point does not take away other potential qualities of R, like its potential to contribute to the rewarding of high quality reviews. So we are left with points 2-4 above. The authors take into account journals' Impact Factor (IF) and want to use a measure of the quality of the individual reviews that is standardized, but it is not entirely clear how they want to achieve this. Also, it is not immediately clear how R deals with the fact that higher IFs are associated with lower word counts in some journals (Science and Nature have clear word limits, whereas some substantive journals with lower IFs have none).

Authors: We agree that the original description of the R-index parameters was insufficient. To make clearer how the R-index will be comparable across journals, we now improved this excerpt by 1) justifying the IF and manuscript length as valid working proxies (L61-69), and 2) detailing the definition of the s-score to make it more objective and standardized across disciplines (L70-86). Usually highly productive researchers in a given area are invited to review for higher rank journals, so we see the IF as a proxy for one's prestige in their field as a reviewer. IFs are inherently different across disciplines and we propose that its square-rooted value can alleviate the differences. There is no single measure for time spent in a review; manuscript length is a very intuitive measure of it. Clearly, word count is a common feature of all manuscripts, despite

inherent variations among reviewers' abilities to review longer or more methodological papers, and variations across disciplines. The rescaled word count (w/10^4) used in the formula reduces the weight of such disparity in the index calculation. Yet, we argue that because journals with higher impact factors usually require shorter manuscripts, we could expect *IF* and *w* balancing each other out in the final index calculation. Finally, our proposed measure of the quality of reviews, the s-score, is intended to rank the reviews in a standardized way (from 0 to 1) based on criteria that we are intrinsic in any review and so useful tools for editors of any journal: punctuality, utility to authors, utility to editors and impact of the review report. We now delineate these criteria better in the L70-86, further suggesting a way to make it more objective and readily comparable across journals of different disciplines.

Goal 3: Fairness across career stages

Reviewer 1*:* As the authors argue, R does appear to deal with point (3), which is probably a good thing given the research showing that early and mid-career scientists write the best reviews and are often (at least in my experience at PLOS ONE) more willing to review.

Authors: We appreciate you highlighting the qualities of the R-index. As early career scientist ourselves, we felt it necessary to deal with what appears to be an unspoken truth about review – as a result a key part of developing R was to ensure that it was fair across career stages. Given reviewer#2's comments about the ratios of career stages and proportions of reviews, we are curious, if Dr. Wicherts would be willing to pers. comm. his experience at PLOS One to further support our model; and if so, we have temporarily included this in L116 and L191 which can be removed if not consented to

Goal 4: Community agreement on parameters

Reviewer 1: Perhaps the most essential question is (4): would all (or at least most) stakeholders agree with the inclusion of w, s, and IF in the index? Here I have some doubts. I could envision a lot of my colleagues arguing vigorously against the adoption of the IF (although I am not as critical as they are regarding IF), and there may also be issues with w. For instance, I also review for technical journals in my field and these reviews typically require much more time (e.g., because I need to check the formulas) despite the fact that the number of words in the manuscripts is limited.

The core question about support for the inclusion of these ingredients is probably mostly an empirical one. Hence, I feel that the current work would be considerably strengthened by adding data on the support for the R index (and its ingredients) in various fields. Similarly, one would need to have more data in order to get a sense of the weighting that the current R index uses for IF, w, and s. The current rationale for the use of the current weights did not entirely convince me and needs more work.

Authors: Dr. Wicherts review highlights a key truth about quantifying science: it has the right combination of appeal and controversy and no one ever agrees with the metrics entirely. There is dissent about the H-index just as there is varying agreement with Impact Factor, and the extent of their usage. Accordingly, this disagreement will exist for any attempt to quantify the reviewers' contributions. As is evident in the reviews for this manuscript in which the two reviewers suggests different improvements for the proposed index. In all fairness, it is unlikely that any metric would completely satisfy the entire scientific community. Despite the general agreement on a reward that is proportional to the number of reviewed manuscripts, we would expect arguments in favour and against differing additional parameters of the index equation. There is no single recipe for such an index. We attempted to formulate one that is as fair as possible, captures as many

aspects as possible of reviewers' contribution and yet still uses a minimum number of working proxies. To support the proposed index formulation in the revised manuscript, we have improved the description and justification each parameter (L56-86) and also provide supplementary analyses modelling the performance of 8 reduced versions of the index (see the updated SM3 and the new figure S4). We suggest that removing parameters only changes the absolute scale of the index, and the original formulation weights the number of reviewed papers with other aspects of contributions through peer-reviewing (L95, L135-138, ESM3).

Nonetheless, we absolutely agree with the reviewer on the need of empirical data to test the index performance beforehand. However, this is not possible at this stage simply because there are no such data available—either because journals do not yet keep track of these parameters we need, or more likely, they are not willing to share such information at this stage. We truly had a hard time gathering the empirical data from the journals and, ultimately we got a database from a single journal. This is why we decided to model realistically-generated data to predict R-index outputs. The greatest advantage of the simulations is the ability of generating large amounts of realistic data. Our first simulations were performed with a very large sample size and a high number of reviewed papers per reviewer (50,000 reviewers and 2,875,000 manuscripts: 27,000 early-career researchers reviewing 75 manuscripts each, 16,000 mid-careers reviewing 40 manuscripts each, and 3,500 opportunist and 3,500 specialist lead-researchers reviewing 30 manuscripts each). To double-check if the sample size was sufficient, during this review process we re-ran all the simulations with 10 times more data (500,000 reviewers and 28,750,000 manuscripts), using the empirical IF distribution with 7,514 journals. This given us the very high average of 3,826 manuscripts per journal. The results were exactly the same shown in the Figures 1 and 2— except, obviously, that the clouds of points in the Figure 1 had 10 times more points and the frequencies of the histograms (y-axis of the Figure 2) were tenfold. This shows the index is reliable and that we have found a consistent pattern. More importantly here, this shows our original sample size is enough to evaluate the weights of R-index parameters and likely to be a realistic sample. For instance, a top journal such as *Nature* receives about 200 manuscripts per week (9600 per year); our original sample of almost 3 million manuscripts would represent the bulk of submissions to at least 300 journals of the same caliber. Given that lower rank journals inherently receive less than 9600 manuscripts per year, we are likely covering a realistic number of manuscripts being reviewed per year.

We believe our metric can be easily implemented but more importantly it will stimulate the debate on the need for quantifying reviewers' contributions, and so provide the impetus for the collection of the empirical data needed. We are actually very excited to test the index as soon as this data is made available. Just like H-index, Impact Factor and other metrics, we expected and hope that R-index evolves as we learn from the real data.

Comments from the Reviewer 2, Dr. David Cameron Duffy

Dear Dr. Duffy,
We are grateful for the thorough and signed review that helped us improve the quality of this work. Please find below how we addressed each of your comments in the new version of the manuscript.

Comment # 1 by Reviewer 2:
The English needs work and the discussion needs to be tightened up.

Response by the Authors: The manuscript was authored by a native English speaker, and reviewed prior to submission by another. While we would be happy to address any grammatical, structural, or spelling mistakes, in addition to any stylistic points brought forward by the reviewer; the vagueness of this comment is a challenge to address specifically. In an attempt to broadly address this point, we have thoroughly revised the final version of the manuscript after improved and reformatted the discussion into a separate section (see comment #9).

Comment # 2 by Reviewer 2:
I think the model can be better explained, with more details.

Response by the Authors: We followed the reviewer's suggestion and expanded the description of all parameters of the index (L56-86).

Comment # 3 by Reviewer 2:
More specifically, is the number of words in a paper really a proxy for time spent? And is this the published number? A reviewer may put a lot of time into helping shorten a paper but their zeal would in fact be penalized. Also some fields are wordier than others.

Response by the Authors: The main idea of our index is to quantify the different facets of contributing as an individual reviewer. Given the time trade-off between publishing and reviewing, we wanted a proxy for time spent in the review. We concluded that the fairest proxy was the manuscript length, which is also a common feature of any manuscript of any discipline. In reality, there is no single measure for time invested, as manuscript lengths vary across disciplines and reviewers vary in their ability to review longer manuscripts or with more mathematical formulae. Although not impeccable, word count is still an effective and intuitive working proxy for the time taken to complete a review that we feel will be accepted widely. As for the distinction been submitted length and final length, we intended on the use of the word count of the submitted manuscript, given that this is the length of the manuscript when reviewed by the reviewer. Furthermore, this is often data collected during the online submission process, thereby making it simple to access for the journals and editors. Furthermore, w is rescaled by 10^4 thereby minimizing the difference between these two values (in the same way it reduces disparities among disciplines). Consider an extreme example: a long manuscript (say 10,000 words) and a reviewer who did the great job of reducing 3 pages of it (300 words/page for a standard Word manuscript with 12pt and double-sized spacing). For the manuscript, the parameter is w=1 and for the paper it is w=0.91 ([10,000 – 900] /10^4), which would represent a reduction of only 9% of the contribution of that single reviewed paper for the reviewer's overall R-index. Nonetheless, we have made it clear in text that the word length is derived from the submitted manuscript (L61-65).

Comment # 4 by Reviewer 2:
For the s-score how does one standardize across journals (and across editors) and how does one weigh the various aspects of a review? Only whether the reviewer met the deadline is objective; the rest are not.

Response by the Authors: We agree with the reviewer: the s-score is a subjective measure to a certain degree. But subjective measures are useful when a standardized objective one is not immediately available or even possible. Our intention was suggesting that the quality of a review could be ranked and evaluated in a standardized way. For that we propose a score ranging from 0 to 1 that will weight each review based on intuitive proxies (impact, usefulness to the editors, thoroughness and time taken). The actual

responsibility of ranking and scoring a paper, however, it is entirely the editor's. We understand that the editors' job is to evaluate the quality of a paper, using the reviewers reports as the most appropriate tools; thus it seem logical that editors would be the only and the best ones to evaluate the quality of their tools. Originally, we have suggested a minimum of proxies that should be taken into consideration, some of which were objective, others were subjective. In the revised version, we greatly improved the description of our proxies and proposed a Multi-step Likert-type scales as a new way of standardizing s-score across disciplines (L70-86). Defining a single way of weighting the aspects of a review across editors of different disciplines would be, quite honestly, almost arrogant of us, and certainly inaccurate. Our suggestions give a framework for the editors, but it is still flexible, as we believe that one should trust the editors' ability to evaluate each review as an individual case.

Comment # 5 by Reviewer 2:
l. 76 early career reviewing "many". But until they are known and published, they will not be asked to be reviewers. This and what follows assumes a Peer of Science arrangement where reviewers know what manuscripts are out there and can chose to do lots of reviews. For many journals, reviewing is by invitation only.

Response by the Authors: To be clear, our index is aimed to the traditional peer-review system that works under invitation only. We concur that it is likely that the opportunities for review are driven by publications: the more one publishes, the more likely you will be invited to review—although there is no available empirical data that confirms this argument either. Our own experience as early-career researchers, as well as that of reviewer 1's experience as an editor at PLOS One, matches with the data available by the online survey we cited in this paragraph: there are a larger population of early-careers and these tend to review more. Furthermore, offers to review are often deflected by senior researchers onto their graduate students and fellows through suggestions of alternative reviewers. Our simulations attempt to mimic the real world by creating career stage categories that reflect the proportions given by the real data. We now clarify this issue adding the proportions to this paragraph (L112-117).

Comment # 6 by Reviewer 2:
l. 80-85 opportunist and specialist--what is the basis for this? Many might pursue a 'mixed' strategy. For example, a senior person might be both opportunistic, if offered the chance to review for Nature or Science, but willing to review lots of papers within their own field.

Response by the Authors: We absolutely agree with the reviewer that other possible strategies are possible, and indeed such mixed strategy is very likely to occur. Empirical data on researchers reviewing habits would be of great help here. But such data is just not available, as surveys on this type of behaviour do not exist. R-index, and particular its ratio with H-index, would begin to elucidate the reality within our community, and differing patterns between fields. For now, we overcame the limitations of real-world data with simulation experiments. Our point when simulating reviewer strategies was to portray extremist habits, which could cover the broad range of possibilities in between and provide insights on the spectrum of R-index's performance. By doing so, we considered that mixed strategies (such as the referred case in which individuals are both opportunistic and specialist) would lay within the range of possibilities considered by the extreme strategies. To follow specifically the reviewer's suggestion, we now modelled two new mixed strategies that basically differ on the number of reviews performed, being either from low- or high rank journals. We have shown that the R-indices of reviewers following either mixed strategy lay within the R-index of opportunist and specialist reviewers, ultimately cross-validating our original approach. The new

results (Figure S2) are available in the new supplementary material SM2 (as well as the simulations are in the R package) and we make reference to it in the main text (L110-111).

Comment # 7 by Reviewer 2:
94 "tendency" within the model's conditions

Response by the Authors: The reviewer is correct. However, such a tendency was only included within the models conditions because of the empirical evidence (given by the reference 12). This sentence now makes this very clear (L143).

Comment # 8 by Reviewer 2:
98 how did you adjust for editor's feedback?

Response by the Authors: We first model the R-index outcomes for all reviewer strategies drawing the editor's feedback (i.e. the s-score) from an empirical distribution made available to use from a real journal's editor. The results are in the Figure 1A. We then adjusted the editor's feedback by assigning s-scores for each reviewer strategy drawing from different ranges of the empirical distribution, i.e. using a stratified sampling of the empirical s-score distribution. We agree with the reviewer that the previous version lack clarity and we now make these details available in the Methods section (L118-126), as well as in the new Supplementary Material SM2.

Comment # 9 by Reviewer 2:
This needs bait more explanation, as in l. 100 why would opportunistic reviewers provide poor reviews to top journals?

Response by the Authors: We agree with the reviewer that the excerpt was unclear. As detailed in the comment#6, our simulations aimed to portray a broad range of possibilities of reviewing habits. While there are no empirical evidences that reviewers for top journals do a poor job, with the implementation of the R-index such strategy could arise as an attempt to game the system by providing quick, poor reviews to boost the index. Our modelling aimed to evaluate how this strategy would perform, in order to assess how easily one could game the R-index. We now made our modelling goals more clear in text L127-130.

Comment # 10 by Reviewer 2:
l. 110 "standing in the field" not clear how it does this

Response by the Authors: We used "standing in the field" in this sentence to mean the inclusion of the journal's impact factor on the index. This is now clearly stated in the index formulation, L67-68. We argue that usually renowned researchers in a given area (i.e. highly productive in that area) are the ones that are invited to higher rank journals. With our own experience as early-career researchers, and of our renowned colleagues', it is less likely that early-career or researchers who publish few papers will get invitations to review for top-journals. Your previous comment suggests that you agree with this statement, in that you suggested that being "known" increases the likelihood of invitations. Similarly, in your *The Scientist* article you stated that, "I believe being asked to referee reflects one's true standing in a field" and so suggested using IF as a multiplier to reflect standing in the field. Therefore, we considered impact factor as a proxy for one's prestige in their field as a reviewer. To remove any possible ambiguities, we now better explain the impact factor in the index formulation (L65-70).

Comment # 11 by Reviewer 2:
I guess I feel uneasy at the mixed results/discussion as it is hard to tell where results end and discussion begins, i.e. l. 111 "unheralded pillars", l. 113 :"in our experience", "we assume", l. 118 "hard working", l. 119 "typically", ;. 121 "daunting", l. 126 "inevitably" l. 144-145 results appear, after a long run of discussion. I'd suggest separating them and elaborating on the results and shortening the discussion.

Response by the Authors: We followed the reviewer's suggestion and attempted to delineate the results from the discussion. All of the quoted text appears now in the discussion.

Comment # 12 by Reviewer 2:
l. 147. judging journals by metrics. Editors will game this, as they will score their own reviewers higher to elevate R (and to avoid annoying their reviewers who will avoid reviewing for journals with tough-scoring editors), unless there are quantitative standards.

Response by the Authors: In truth, this practice would only result in extra work for the editors and a decrease in the quality and utility of reviews overall. If editors score poor reviews highly in an attempt to game the journal's R-index, and editors (within or between journals) select reviewers based on high individual R scores, then it will lead to more poor reviewers and more work for the editors. There would be a negative feedback loop which would prevent the practice. As for reviewers refusing reviews from tough-scoring editors, this can be viewed in the opposite framing as well. One could only know that an editor gives consistent low s-scores if they have reviewed for that editor before (unless the journals also publically publish individual editors mean s-scores – which should be encouraged for transparency, but unlikely initially). It's unlikely that editors would invite the reviewer again if they gave them poor scores previously, and therefore the reviewer would not be able to avoid them.

Comment # 13 by Reviewer 2:
l. 156 "practical aid in the publication process". One complaint of reviewers and editors is that sometimes authors expect them to do the work of honing the manuscript for publication. In other words, an author is seen as "just sending it in to see how the reviewers respond", rather than having colleagues vet the paper before submission.

Response by the Authors: That is, in our experience, very true. Using the peer-review as sheer step in the process of publication is one way authors can game the peer-review system. It is, however, a problem with the system itself rather than the proposed index. Given the large amount of manuscripts produced, the rejection probability is already high just because journals cannot accommodate all manuscripts rather than strictly for the quality of the work. So using reviewers from prestigious journal to improve the manuscript for another journal became a fairly common strategy. What we meant by "practical aid in the publication process" is not quite this game. We understand that the key contributions of reviewers' are to judge the merit, and to help improving the quality of the manuscript. Thus, competent reviewers should be able to distinguish between a final product and a work in progress. While in the former case reviewers can indeed improve the work, in the latter, we think, the manuscript should be rejected right away.

Comment # 14 by Reviewer 2:
Given the gaming that is going on with less reputable on-line journals, the authors should look into how such journals might be used to game R, as they are used to game H and other indices. Indeed they might tinker with the model and see what vulnerabilities it has.

Response by the Authors: That is a very valid point. Even as a debatable evaluation of a journal's reputation, the IF would control for attempts to boost R-index, because the referred "less reputable" online journals have very low, if any, impact factor. (We understand "less reputable online journals" as the bloom journals that rise and fall each month somewhere in the world with tricky similar names to renowned journals). On the reviewer's side, reviewing for such journals would not weight much in their individual index. On the journal's side, the only way editors could try to game R-index would be by overprizing their reviewers with a disproportionally high s-score to advertise the journal's review quality. Still, high s-scores with very low IFs would not considerably lift the journal's average R-index. We tinkered with the simulation by varying the other parameters but fixing IF to a very low number, say 0.001, to show that the index would still be proportionally driven by the number of reviews performed and mainly weighted by the quality of the review (s). These new simulations suggesting that R-index is not vulnerable to such potential game are now available in the electronic supplementary material SM4 and referred in the main results section (L97, L156-158). Please note that while the R-index outputs behave similarly with empirical IF and the very low and fixed IFs, the range of the R-index is largely affected. We believe R-index is robust to such game since the journal's Impact Factor is accounted for.

Comment # 15 by Reviewer 2:
I do wish the authors would have actually cited my paper rather than referring to it in the acknowledgements, but I am happy in any event that they are advancing the idea.

Response by the Authors: We found that clearly acknowledging how your article has inspired our work would be more appropriate than just citing it, almost anonymously, in the numbered reference list. In the current version, we unmistakably recognize the influence of your work by doing both (L41, L273-274, L313).