

# frontiers RESEARCH TOPICS

## BEYOND OPEN ACCESS: VISIONS FOR OPEN EVALUATION OF SCIENTIFIC PAPERS BY POST-PUBLICATION PEER REVIEW

Hosted by  
Nikolaus Kriegeskorte and Diana Deca



**frontiers in**  
**COMPUTATIONAL NEUROSCIENCE**



# frontiers

## FRONTIERS COPYRIGHT STATEMENT

© Copyright 2007-2012  
Frontiers Media SA.  
All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, as well as all content on this site is the exclusive property of Frontiers. Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Articles and other user-contributed materials may be downloaded and reproduced subject to any copyright or other notices. No financial payment or reward may be given for any such reproduction except to the author(s) of the article concerned.

As author or other contributor you grant permission to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

Cover image provided by Ibbl sarl, Lausanne CH

ISSN 1664-8714

ISBN 978-2-88919-068-3

DOI 10.3389/978-2-88919-068-3

## ABOUT FRONTIERS

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

## FRONTIERS JOURNAL SERIES

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing.

All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

## DEDICATION TO QUALITY

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.

By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

## WHAT ARE FRONTIERS RESEARCH TOPICS?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area!

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: [researchtopics@frontiersin.org](mailto:researchtopics@frontiersin.org)

# BEYOND OPEN ACCESS: VISIONS FOR OPEN EVALUATION OF SCIENTIFIC PAPERS BY POST-PUBLICATION PEER REVIEW

Hosted By:

**Nikolaus Kriegeskorte**, Medical Research Council Cognition and Brain Sciences Unit, United Kingdom

**Diana Deca**, TUM, Germany



A scientific publication system needs to provide two basic services: access and evaluation. The traditional publication system restricts the access to papers by requiring payment, and it restricts the evaluation of papers by relying on just 2-4 pre-publication peer reviews

and by keeping the reviews secret. As a result, the current system suffers from a lack of quality and transparency of the peer-review evaluation process, and the only immediately available indication of a new paper's quality is the prestige of the journal it appeared in.

*Open access* is now widely accepted as desirable and is slowly beginning to become a reality. However, the second essential element, evaluation, has received less attention. *Open evaluation*, an ongoing post-publication process of transparent peer review and rating of papers, promises to address the problems of the current system. However, it is unclear how exactly such a system should be designed.

The evaluation system steers the attention of the scientific community and, thus, the very course of science. For better or worse, the most visible papers determine the direction of each field and guide funding and public policy decisions. Evaluation, therefore, is at the heart of the entire endeavor of science. As the number of scientific publications explodes, evaluation and selection will only gain importance. A grand challenge of our time, therefore, is to design the future system, by which we evaluate papers and decide which ones deserve broad attention.

So far scientists have left the design of the evaluation process to journals and publishing companies. However, the steering mechanism of science should be designed by scientists. The cognitive, computational, and brain sciences are best prepared to take on this task, which will involve social and psychological considerations, software design, and modeling of the network of scientific papers and their interrelationships.

This Research Topic in *Frontiers in Computational Neuroscience* collects *visions for a future system of open evaluation*. Because critical arguments about the current system abound, these papers will focus on constructive ideas and comprehensive designs for open evaluation systems. Design decisions include: Should the reviews and ratings be entirely transparent, or should some aspects be kept secret? Should other information, such as paper downloads be included in the evaluation? How can scientific objectivity be strengthened and political motivations weakened in the future system? Should the system include signed and authenticated reviews and ratings? Should the evaluation be an ongoing process, such that promising papers are more deeply evaluated? How can we bring science and statistics to the evaluation process (e.g. should rating averages come with error bars)? How should the evaluative information about each paper (e.g. peer ratings) be combined to prioritize the literature? Should different individuals and organizations be able to define their own evaluation formulae (e.g. weighting ratings according to different criteria)? How can we efficiently transition toward the future system?

Ideally, the future system will derive its authority from a scientific literature on community-based open evaluation. We hope that these papers will provide a starting point.



# Table of Contents

- 06    *An Emerging Consensus for Open Evaluation: 18 Visions for the Future of Scientific Publishing***  
Nikolaus Kriegeskorte, Alexander Walther and Diana Deca
- 11    *Nine Criteria for a Measure of Scientific Output***  
Gabriel Kreiman and John Maunsell
- 17    *Toward a New Model of Scientific Publishing: Discussion and a Proposal***  
Dwight J. Kravitz and Chris I. Baker
- 29    *Alternatives to Peer Review: Novel Approaches for Research Evaluation***  
Aliaksandr Birukou, Joseph Rushton Wakeling, Claudio Bartolini, Fabio Casati, Maurizio Marchese, Katsiaryna Mirylenka, Nardine Osman, Azzurra Ragone, Carles Sierra and Aalam Wassef
- 41    *Fair and Open Evaluation May Call for Temporarily Hidden Authorship, Caution When Counting the Votes, and Transparency of the Full Pre-publication Procedure***  
Talis Bachmann
- 44    *Open Peer Review by a Selected-Papers Network***  
Christopher Lee
- 59    *Maintaining Live Discussion in Two-Stage Open Peer Review***  
Erik Sandewall
- 70    *Tracking Replicability as a Method of Post-Publication Open Evaluation***  
Joshua K. Hartshorne and Adena Schachner
- 84    *Network-based Statistics for a Community Driven Transparent Publication Process***  
Jan Zimmermann, Alard Roebroek, Kamil Uludag, Alexander T. Sack, Elia Formisano, Bernadette Jansma, Peter De Weerd and Rainer Goebel
- 89    *Letting the Daylight in: Reviewing the Reviewers and Other Ways to Maximize Transparency in Science***  
Jelte M. Wicherts, Rogier A. Kievit, Marjan Bakker and Denny Borsboom
- 98    *Decoupling the Scholarly Journal***  
Jason Priem and Bradley M. Hemminger
- 111    *Learning from Open Source Software Projects to Improve Scientific Review***  
Satrajit S. Ghosh, Arno Klein, Brian Avants and K. Jarrod Millman
- 122    *Aggregating Post-publication Peer Reviews and Ratings***  
Răzvan V. Florian
- 130    *FOSE: a Framework for Open Science Evaluation***  
Alexander Walther and Jasper J. F. van den Bosch

- 138** *Multi-Stage Open Peer Review: Scientific Evaluation Integrating the Strengths of Traditional Peer Review with the Virtues of Transparency and Self-Regulation*  
Ulrich Pöschl
- 154** *The Evaluation of Research Papers in the XXI Century. The Open Peer Discussion System of the World Economics Association*  
Grazialetto-Gillies
- 161** *Post-Publication Peer Review: Opening Up Scientific Conversation*  
Jane Hunter
- 163** *Designing Next-generation Platforms for Evaluating Scientific Output: What Scientists can Learn from the Social Web*  
Tal Yarkoni
- 176** *Open Evaluation: A Vision for Entirely Transparent Post-Publication Peer Review and Rating for Science*  
Nikolaus Kriegeskorte



# An emerging consensus for open evaluation: 18 visions for the future of scientific publishing

Nikolaus Kriegeskorte<sup>1\*</sup>, Alexander Walther<sup>1</sup> and Diana Deca<sup>2</sup>

<sup>1</sup> Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK

<sup>2</sup> Institute of Neuroscience, Technische Universität München, Munich, Germany

\*Correspondence: nikokriegeskorte@gmail.com

## Edited by:

Misha Tsodyks, Weizmann Institute of Science, Israel

## Reviewed by:

Misha Tsodyks, Weizmann Institute of Science, Israel

A scientific publication system needs to provide two basic services: access and evaluation. The traditional publication system restricts the access to papers by requiring payment, and it restricts the evaluation of papers by relying on just 2–4 pre-publication peer reviews and by keeping the reviews secret. As a result, the current system suffers from a lack of quality and transparency of the peer review process, and the only immediately available indication of a new paper's quality is the prestige of the journal it appeared in.

Open access (OA) is now widely accepted as desirable and is beginning to become a reality. However, the second essential element, evaluation, has received less attention. Open evaluation (OE), an ongoing post-publication process of transparent peer review and rating of papers, promises to address the problems of the current system and bring scientific publishing into the twenty-first century.

Evaluation steers the attention of the scientific community, and thus the very course of science. For better or worse, the most visible papers determine the direction of each field, and guide funding and public policy decisions. Evaluation, therefore, is at the heart of the entire endeavor of science. As the number of scientific publications explodes, evaluation, and selection will only gain importance. A grand challenge of our time, therefore, is to design the future system, by which we evaluate papers and decide which ones deserve broad attention and deep reading. However, it is unclear how exactly OE and the future system for scientific publishing should work. This motivated us to edit the Research Topic “Beyond open access: visions for open evaluation of scientific papers by post-publication peer review” in Frontiers in Computational Neuroscience. The Research Topic includes 18 papers, each going beyond mere criticism of the status quo and laying out a detailed vision for the ideal future system. The authors are from a wide variety of disciplines, including neuroscience, psychology, computer science, artificial intelligence, medicine, molecular biology, chemistry, and economics.

The proposals could easily have turned out to contradict each other, with some authors favoring solutions that others advise against. However, our contributors' visions are largely compatible. While each paper elaborates on particular challenges, the solutions proposed have much overlap, and where distinct solutions are proposed, these are generally compatible. This puts us in a position to present our synopsis here as a coherent

blueprint for the future system that reflects the consensus among the contributors.<sup>1</sup> Each section heading below refers to a design feature of the future system that was a prevalent theme in the collection. If the feature was overwhelmingly endorsed, the section heading below is phrased as a statement. If at least two papers strongly advised against the feature, the section heading is phrased as a question. **Figure 1** visualizes to what extent each paper encourages or discourages the inclusion of each design feature in the future system. The ratings used in **Figure 1** have been agreed upon with the authors of the original papers.<sup>2</sup>

## SYNOPSIS OF THE EMERGING CONSENSUS

### THE EVALUATION PROCESS IS TOTALLY TRANSPARENT

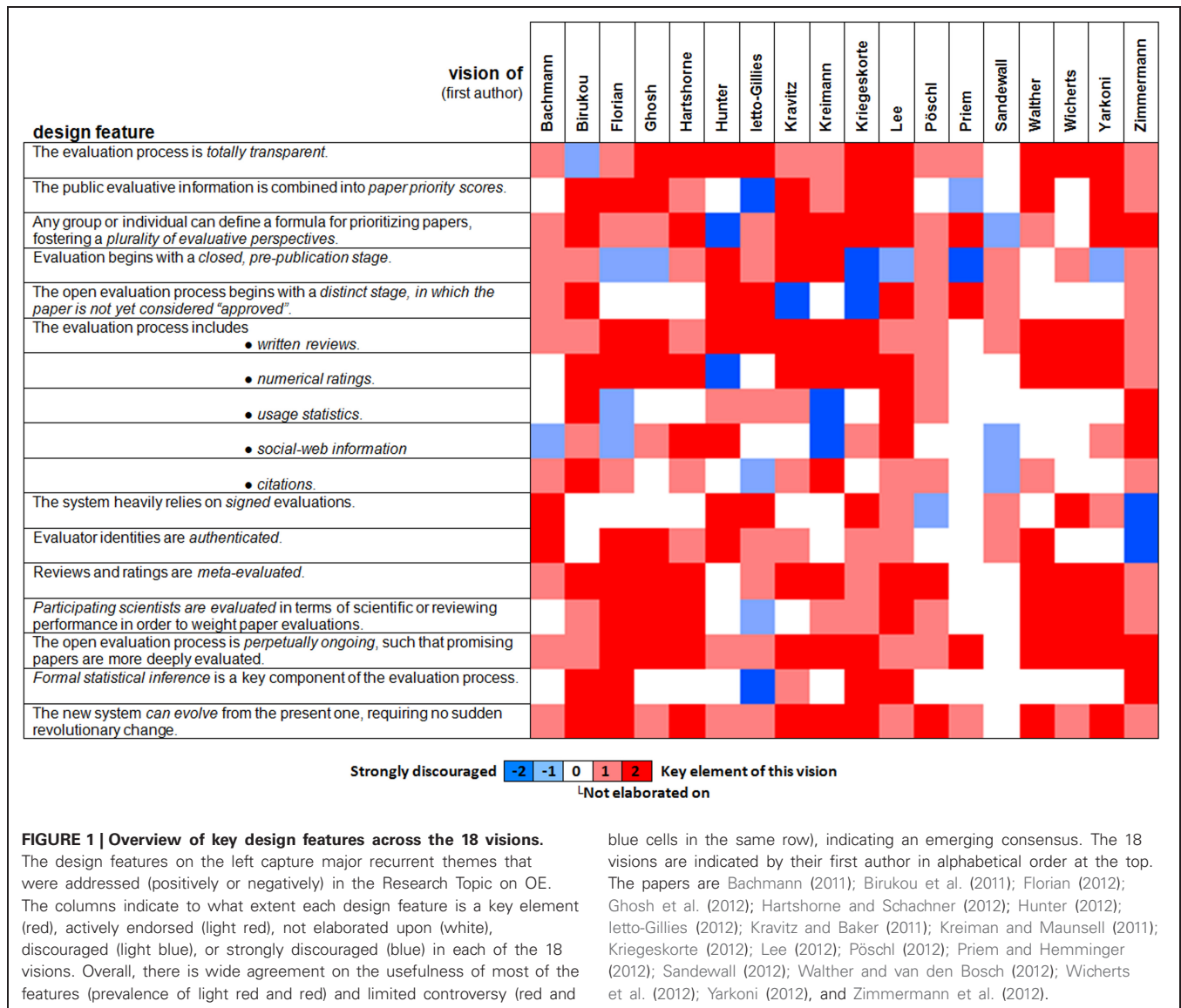
Almost all of the 18 visions favor *total transparency*. Total transparency means that all reviews and ratings are instantly published. This is in contrast to current practice, where the community is excluded and reviews are initially only visible to editors and later on to the authors (and ratings are often only visible to editors). Such secrecy opens the door to self-serving reviewer behavior, especially when the judgments are inherently subjective, such as the judgment of the overall significance of a paper. In a secret reviewing system, the question of a paper's significance may translate in some reviewers' minds to the question “How comfortable am I with this paper gaining high visibility now?” In a transparent evaluation system, the reviews and reviewers are subject to public scrutiny, and reviewers are thus more likely to ask themselves the more appropriate question “How likely is it that this paper will ultimately turn out to be important?”

### THE PUBLIC EVALUATIVE INFORMATION IS COMBINED INTO PAPER PRIORITY SCORES

In a totally transparent evaluation process, the evaluative information (including reviews and ratings) is publicly available.

<sup>1</sup>The consensus, of course, is only among the contributors to this collection. A consensus among the scientific community at large has yet to be established. Note that scientists critical of the general idea of OE would not have chosen to contribute here. Nevertheless, assuming OE is seen as desirable, the collection does suggest that independent minds will produce compatible visions for how to implement it.

<sup>2</sup>With the exception of Erik Sandewall, whom we could not reach before this piece went to press.



Most of the authors suggest the use of functions that combine the evaluative evidence into an overall *paper priority score* that produces a ranking of all papers. Such a score could be computed as an average of the ratings. The individual ratings could be weighted in the average, so as to control the relative influence of different rating scales (e.g., reliability vs. novelty vs. importance of the claims) and to give greater weight to raters that are either highly regarded in the field (by some quantitative measure, such as the h-index) or have proved to be reliable raters in the past.

#### ANY GROUP OR INDIVIDUAL CAN DEFINE A FORMULA FOR PRIORITIZING PAPERS, FOSTERING A PLURALITY OF EVALUATIVE PERSPECTIVES

Most authors support the idea that a *plurality of evaluative perspectives* on the literature is desirable. Rather than creating a centralized black-box system that ranks

the entire literature, any group or individual should be enabled to access the evaluative information and combine it by an arbitrary formula to prioritize the literature. A constant evolution of competing priority scores will also make it harder to manipulate the perceived importance of a paper.

#### SHOULD EVALUATION BEGIN WITH A CLOSED, PRE-PUBLICATION STAGE?

Whether a *closed, pre-publication stage* of evaluation (such as the current system's secret peer review) is desirable is controversial. On the one hand, the absence of any pre-publication filtering may open the gates to a flood of low-quality publications. On the other hand, providing permanent public access to a wide range of papers, including those that do not initially meet enthusiasm, may be a strength rather than a weakness. Much brilliant science was initially misunderstood. Pre-publication filtering comes at

the cost of a permanent loss of value through errors in the initial evaluations. The benefit of publishing all papers may, thus, outweigh the cost of providing the necessary storage and access. “Publish, then filter” is one of the central principles that lend the web its power (Shirky, 2008). It might work equally well in science as it does in other domains, with *post-publication* filtering preventing the flood from cluttering our view of the literature.

#### **SHOULD THE OPEN EVALUATION BEGIN WITH A DISTINCT STAGE, IN WHICH THE PAPER IS NOT YET CONSIDERED “APPROVED”?**

Instead of a closed, pre-publication evaluation, we could define a *distinct initial stage of the post-publication open evaluation* that determines whether a paper receives an “approved” label. Whether this is desirable is controversial among the 18 visions. One argument in favor of an “approved” label is that it could serve the function of the current notion of “peer reviewed science,” suggesting that the claims made are somewhat reliable. However, the strength of post-publication OE is ongoing and continuous evaluation. An “approved” label would create an artificial dichotomy based on an arbitrary threshold (on some paper evaluation function). It might make it more difficult for the system to correct its errors as more evaluative evidence comes in (unless papers can cross back over to the “unapproved” state). Another argument in favor of an initial distinct stage of OE is that it could serve to incorporate an early round of review and revision. The authors could choose to either accept the initial evaluation, or revise the paper and trigger re-evaluation. However, revision and re-evaluation would be possible at any point of an open evaluation process anyway. Moreover, authors can always seek informal feedback (either privately among trusted associates or publicly via blogs) prior to formal publication.

#### **THE EVALUATION PROCESS INCLUDES WRITTEN REVIEWS, NUMERICAL RATINGS, USAGE STATISTICS, SOCIAL-WEB INFORMATION, AND CITATIONS**

There is a strong consensus that the OE process should include *written reviews* and *numerical ratings*. These classical elements of peer review continue to be useful. They represent explicit expert judgments and serve an important function that is distinct from the function of *usage statistics* and *social-web information*, which are also seen as useful by some of the authors. In contrast to explicit expert judgments, usage statistics, and social-web information may highlight anything that receives attention (of the positive or negative variety), thus potentially valuing buzz and controversy over high-quality science. Finally, *citations* provide a slow signal of paper quality, emerging years after publication. Because citations are slow to emerge, they cannot replace the other signals. However, they arguably provide the ultimately definitive signal of a paper’s de-facto importance.

#### **THE SYSTEM UTILIZES SIGNED (ALONG WITH UNSIGNED) EVALUATIONS**

*Signed evaluations* are a key element of five of the visions, only one vision strongly discourages heavy reliance on signed evaluations.

When an evaluation is signed, it affects the evaluator’s reputation. High-quality signed evaluations can help build a scientist’s reputation (thus motivating scientists to contribute). Conversely, low-quality signed evaluations can hurt a scientist’s reputation (thus motivating high standards in rating and reviewing). Signing creates an incentive for objectivity and a disincentive for self-serving judgments. But as signing adds weight to the act of evaluation, it might also create hesitation. Hesitation to provide a rash judgment may be desirable, but the system does require sufficient participation. Moreover, signing may create a disincentive to present critical arguments as evaluators may fear potential social consequences of their criticism. The OE system should therefore collect both signed and unsigned evaluations, and combine the advantages of these two types of evaluation.

#### **EVALUATORS’ IDENTITIES ARE AUTHENTICATED**

*Authentication of evaluator identities* is a key element of five of the visions, one vision strongly discourages it. Authentication could be achieved by requiring login with a password before submitting evaluations. Authenticating the evaluator’s identity does not mean that the evaluator has to publicly sign the evaluation, but would enable the system to exclude lay people from the evaluation process and to relate multiple reviews and ratings provided by the same person. This could be useful for assessing biases and estimating the predictive power of the evaluations. Arguments against authenticating evaluator identities (unless the evaluator chooses to sign) are that it creates a barrier to participation and compromises transparency (the “system,” but not the public knows the identity). However, authentication could use public aliases, allowing virtual evaluator identities (similar to blogger identities) to be tracked without any secret identity tracking. Note that (1) anonymous, (2) authenticated-unsigned, and (3) authenticated-signed evaluations each have different strengths and weaknesses and could all be collected in the same system. It would then fall to the designers of paper evaluation functions to decide how to optimally combine the different qualities of evaluative evidence.

#### **REVIEWS AND RATINGS ARE META-EVALUATED**

Most authors suggest *meta-evaluation of individual evaluations*. One model for meta-evaluation is to treat reviews and ratings like papers, such that paper evaluations and meta-evaluations can utilize the same system. Paper evaluation functions could retrieve meta-evaluations recursively and use this information for weighting the primary evaluations of each paper. None of the contributors to the Research Topic object to meta-evaluation.

#### **PARTICIPATING SCIENTISTS ARE EVALUATED IN TERMS OF SCIENTIFIC OR REVIEWING PERFORMANCE IN ORDER TO WEIGHT PAPER EVALUATIONS**

Almost all authors suggest that the system *evaluate the evaluators*. Evaluations of evaluators would be useful for weighting the multiple evaluations a given new paper receives. Note that this will require some form of authentication



of the evaluators' identities. Scientists could be evaluated by combining the evaluations of their publications. A citation-based example of this is the h-index, but the more rapidly available paper evaluations provided by the new system could also be used to evaluate an individual's scientific performance. Moreover, the predictive power of a scientist's previous evaluations could be estimated as an index of reviewing performance. An evaluation might be considered predictive to the extent that it deviates from previous evaluations, but matches later aggregate opinion.

#### THE OPEN EVALUATION PROCESS IS PERPETUALLY ONGOING, SUCH THAT PROMISING PAPERS ARE MORE DEEPLY EVALUATED

Almost all authors suggest a *perpetually ongoing* OE process. Ongoing evaluation means that there is no time limit on the evaluation process for a given paper. This enables the OE process to accumulate deeper and broader evaluative evidence for promising papers, and to self-correct when necessary, even if the error is only discovered long after publication. Initially exciting papers that turn out to be incorrect could be debunked. Conversely, initially misunderstood papers could receive their due respect when the field comes to appreciate their contribution. None of the authors objects to perpetually ongoing evaluation.

#### FORMAL STATISTICAL INFERENCE IS A KEY COMPONENT OF THE EVALUATION PROCESS

Many of the authors suggest a role for *formal statistical inference in the evaluation process*. Confidence intervals on evaluations would improve the way we allocate our attention, preventing us from preferring papers that are not significantly preferable and enabling us to appreciate the full range of excellent contributions, rather than only those that find their way onto a stage of limited size, such as the pages of *Science* and *Nature*. To the extent that excellent papers do not significantly differ in

their evaluations, the necessary selection would rely on content relevance.

#### THE NEW SYSTEM CAN EVOLVE FROM THE PRESENT ONE, REQUIRING NO SUDDEN REVOLUTIONARY CHANGE

Almost all authors suggest that *the ideal system for scientific publishing can evolve* from the present one, requiring no sudden revolutionary change. The key missing element is a powerful general OE system. An OE system could initially serve to more broadly and deeply evaluate papers published in the current system. Once OE has proven its power and its evaluations are widely trusted, traditional pre-publication peer review will no longer be needed to establish a paper as part of the literature. Although the ideal system can evolve, it might take a major public investment (comparable to the establishment of PubMed) to provide a truly transparent, widely trusted OE system that is independent of the for-profit publishing industry.

#### CONCLUDING REMARKS

OA and OE are the two complementary elements that will bring scientific publishing into the twenty-first century. So far scientists have left the design of the evaluation process to journals and publishing companies. However, the steering mechanism of science should be designed by scientists. The cognitive, computational, and brain sciences are best prepared to take on this task, which will involve social and psychological considerations, software design, modeling of the network of scientific papers and their interrelationships, and inference on the reliability and importance of scientific claims. Ideally, the future system will derive its authority from a scientific literature on OE and on methods for inference from the public evaluative evidence. We hope that the largely converging and compatible arguments in the papers of the present collection will provide a starting point.

#### REFERENCES

- Bachmann, T. (2011). Fair and open evaluation may call for temporarily hidden authorship, caution when counting the votes, and transparency of the full pre-publication procedure. *Front. Comput. Neurosci.* 5:61. doi: 10.3389/fncom.2011.00061
- Birukou, A., Wakeling, J. R., Bartolini, C., Casati, F., Marchese, M., Milylenka, K., et al. (2011). Alternatives to peer review: novel approaches for research evaluation. *Front. Comput. Neurosci.* 5:56. doi: 10.3389/fncom.2011.00056
- Florian, R. V. (2012). Aggregating post-publication peer reviews and ratings. *Front. Comput. Neurosci.* 6:31. doi: 10.3389/fncom.2012.00031
- Ghosh, S. S., Klein, A., Avants, B., and Millman, K. J. (2012). Learning from open source software projects to improve scientific review. *Front. Comput. Neurosci.* 6:18. doi: 10.3389/fncom.2012.00018
- Hartshorne, J. K., and Schachner, A. (2012). Tracking replicability as a method of post-publication open evaluation. *Front. Comput. Neurosci.* 6:8. doi: 10.3389/fncom.2012.00008
- Hunter, J. (2012). Post-publication peer review: opening up scientific conversation. *Front. Comput. Neurosci.* 6:63. doi: 10.3389/fncom.2012.00063
- Ietto-Gillies, G. (2012). The evaluation of research papers in the XXI century. The Open Peer Discussion system of the World Economics Association. *Front. Comput. Neurosci.* 6:54. doi: 10.3389/fncom.2012.00054
- Kravitz, D. J., and Baker, C. I. (2011). Toward a new model of scientific publishing: discussion and a proposal. *Front. Comput. Neurosci.* 5:55. doi: 10.3389/fncom.2011.00055
- Kreiman, G., and Maunsell, J. (2011). Nine criteria for a measure of scientific output. *Front. Comput. Neurosci.* 5:48. doi: 10.3389/fncom.2011.00048
- Kriegeskorte, N. (2012). Open evaluation: a vision for entirely transparent post-publication peer review and rating for science. *Front. Comput. Neurosci.* 6:79. doi: 10.3389/fncom.2012.00079
- Lee, C. (2012). Open peer review by a selected-papers network. *Front. Comput. Neurosci.* 6:1. doi: 10.3389/fncom.2012.00001
- Pöschl, U. (2012). Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation. *Front. Comput. Neurosci.* 6:33. doi: 10.3389/fncom.2012.00033
- Priem, J., and Hemminger, B. M. (2012). Decoupling the scholarly journal. *Front. Comput. Neurosci.* 6:19. doi: 10.3389/fncom.2012.00019
- Sandewall, E. (2012). Maintaining live discussion in two-stage open peer review. *Front. Comput. Neurosci.* 6:9. doi: 10.3389/fncom.2012.00009
- Shirky, C. (2008). *Here Comes Everybody: The Power of Organizing Without Organizations*. New York, NY: Penguin Press.
- Walther, A., and van den Bosch, J. F. (2012). FOSE: a framework for open science evaluation. *Front. Comput. Neurosci.* 6:32. doi: 10.3389/fncom.2012.00032
- Wichert, J. M., Kievit, R. A., Bakker, M., and Borsboom, D. (2012). Letting the daylight in: Reviewing the reviewers and other ways to maximize transparency in science.

- Front. Comput. Neurosci.* 6:20. doi: 10.3389/fncom.2012.00020
- Yarkoni, T. (2012). Designing next-generation platforms for evaluating scientific output: what scientists can learn from the social web. *Front. Comput. Neurosci.* 6:72. doi: 10.3389/fncom.2012.00072
- Zimmermann, J., Roebroek, A., Uludag, K., Sack, A. T., Formisano, E., Jansma, B., et al. (2012). Network-based statistics for a community driven transparent publication process. *Front. Comput. Neurosci.* 6:11. doi: 10.3389/fncom.2012.00011
- Received: 23 October 2012; accepted: 24 October 2012; published online: 15 November 2012.
- Citation: Kriegeskorte N, Walther A and Deca D (2012) An emerging consensus for open evaluation: 18 visions for the future of scientific publishing. *Front. Comput. Neurosci.* 6:94. doi: 10.3389/fncom.2012.00094
- Copyright © 2012 Kriegeskorte, Walther and Deca. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



# Nine criteria for a measure of scientific output

Gabriel Kreiman<sup>1,2 \*</sup> and John H. R. Maunsell<sup>3</sup>

<sup>1</sup> Department of Ophthalmology, Children's Hospital, Harvard Medical School, Boston, MA, USA

<sup>2</sup> Department of Neurology, Children's Hospital, Harvard Medical School, Boston, MA, USA

<sup>3</sup> Department of Neurobiology, Harvard Medical School, Boston, MA, USA

## Edited by:

Nikolaus Kriegeskorte, Medical Research Council Cognition and Brain Sciences Unit, UK

## Reviewed by:

Yasser Roudi, Norwegian University of Science and Technology, Norway  
Konrad Koering, Northwestern University, USA

## \*Correspondence:

Gabriel Kreiman, Department of Ophthalmology, Children's Hospital, Harvard Medical School, 1 Blackfan Circle, Karp 11217, Boston, MA 02115, USA.

e-mail: gabriel.kreiman@tch.harvard.edu

Scientific research produces new knowledge, technologies, and clinical treatments that can lead to enormous returns. Often, the path from basic research to new paradigms and direct impact on society takes time. Precise quantification of scientific output in the short-term is not an easy task but is critical for evaluating scientists, laboratories, departments, and institutions. While there have been attempts to quantifying scientific output, we argue that current methods are not ideal and suffer from solvable difficulties. Here we propose criteria that a metric should have to be considered a good index of scientific output. Specifically, we argue that such an index should be quantitative, based on robust data, rapidly updated and retrospective, presented with confidence intervals, normalized by number of contributors, career stage and discipline, impractical to manipulate, and focused on quality over quantity. Such an index should be validated through empirical testing. The purpose of quantitatively evaluating scientific output is not to replace careful, rigorous review by experts but rather to complement those efforts. Because it has the potential to greatly influence the efficiency of scientific research, we have a duty to reflect upon and implement novel and rigorous ways of evaluating scientific output. The criteria proposed here provide initial steps toward the systematic development and validation of a metric to evaluate scientific output.

**Keywords: impact factors, peer review, productivity, scientific output, citation, bibliometric analysis, quality versus quantity, impact**

## INTRODUCTION

Productivity is the ratio of some output value to some input value. In some enterprises productivity can be measured with high precision. A factory can easily measure how many widgets are produced per man-hour of labor. Evaluating scientific productivity, however, is trickier. The input value for scientific productivity is tractable: it might be measured in terms of years of effort by a scientist, research team, department or program, or perhaps in terms of research dollars. It is the output value for scientific productivity that is problematic.

Scientific research produces new knowledge, some fraction of which can lead to enormous returns. In the long run, science evaluates itself. History has a particularly rigorous way of revealing the value of different scientific theories and efforts. Good science leads to novel ideas and changes the way we interpret physical phenomena and the world around us. Good science influences the direction of science itself, and the development of new technologies and social policies. Poor science leads to dead ends, either because it fails to advance understanding in useful ways or because it contains important errors. Poor science produces papers that can eventually feed the fireplace, or in a more modern and ecologically friendly version, the accumulation of electronic documents.

The process of science evaluating itself is slow. Meanwhile, we need more immediate ways of evaluating scientific output. Sorting out which scientists and research directions are currently providing the most useful output is a thorny problem, but it must be done. Scientists must be evaluated for hiring and promotion,

and informed decisions need to be made about how to distribute research funding. The need for evaluation goes beyond the level of individuals. It is often important to evaluate the scientific output of groups of scientists such as laboratories, departments, centers, whole institutions, and perhaps even entire fields. Similarly, funding organizations and agencies need to evaluate the output from various initiatives and funding mechanisms.

Scientific output has traditionally been assessed using peer review in the form of evaluations from a handful of experts. Expert reviewers can evaluate the rigor, value and beauty of new findings, and gauge how they advance the field. Such peer-review constitutes an important approach to evaluating scientific output and it will continue to play a critical role in many forms of evaluation. However, peer review is limited by its subjective nature and the difficulty of obtaining comments from experts that are thorough and thoughtful, and whose comments can be compared across different evaluations. These limitations have driven institutions and agencies to seek more quantitative measures that can complement and sometimes extend thorough evaluation by peers.

In the absence of good quantitative measures of scientific output, many have settled for poor ones. For example, it is often assumed, explicitly, or implicitly, that a long list of publications indicates good output. Using the number of publications as a metric emphasizes quantity rather than quality, when it is the latter that is almost always the value of interest (Siegel and Baveye, 2010; Refinetti, 2011). In an attempt to measure something closer to quality, many turn to journal impact factors (Garfield,



2006). The misuse of journal impact factors in evaluating scientific output has been discussed many times (e.g., Hecht et al., 1998; Amin and Mabe, 2000; Skorka, 2003; Hirsch, 2005; Editors, 2006; Alberts et al., 2008; Castelnovo, 2008; Petsko, 2008; Simons, 2008; Bollen et al., 2009; Dimitrov et al., 2010; Hughes et al., 2010 among many others). We will not repeat the problems with using the impact factors of *journals* to evaluate the output of *individual scientists* here, nor will we focus on the negative effects this use has had on the process of publishing scientific articles. Instead, we note that the persistent misuse of impact factors in the face of clear evidence of its inadequacies must reflect desperation for a quantitative measure of scientific output.

Many measures of scientific output have been devised or discussed. Because most scientific output takes the form of publication in peer-reviewed journals, these measures focus on articles and citations (Bollen et al., 2009). They include a broad range of approaches, such as total number of citations, journal impact factors (Garfield, 2006), *h*-factor (Hirsch, 2005), page ranks, article download statistics, and comments using social media (e.g., Mandavilli, 2011). While all these approaches have merit, we believe that no existing method captures all the criteria that are needed for a rigorous and comprehensive measure of scientific output. Here we discuss what we consider necessary (but not necessarily sufficient) criteria for a metric or index of scientific output. The goal of developing quantitative criteria to evaluate scientific output is not to replace examination by expert reviewers but rather to complement peer-review efforts. The criteria that we propose are aimed toward developing a quantitative metric that is appropriately normalized, emphasizes the quality of scientific output, and can be used for rigorous, reliable comparisons. We do not propose a specific measure, which should be based on extensive testing and comparison of candidate approaches, together with feedback from interested parties. Nevertheless, we believe that a discussion of properties that would make a suitable measure may help progress toward this goal.

We propose that a good index of scientific output will need to have nine characteristics.

## DATA QUALITY AND PRESENTATION

### QUANTITATIVE

Perhaps the most important requirement of a good measure of scientific output is that it be quantitative. The primary alternative, subjective ratings by experts will continue to be important for evaluations, but nevertheless suffers from some important limitations. Ratings by a handful of invited peers, as is normally used in hiring and promoting of scientists, provide ratings of undetermined precision. Moreover, the peers providing detailed comments on different job candidates or grant applications are typically non-overlapping, making it difficult to directly compare their comments.

A further problem with subjective comments is that they put considerable demands on reviewers' time. This makes it impractical to overcome uncertainties about comparisons between different reviewers by reaching out to a very large pool of reviewers for detailed comments. The alternative of getting brief comments from a very large pool of reviewers is also unlikely to work. Several initiatives provide frameworks for peer commentary from large

sets of commenters. Most online journals provide rapid publication of comments from readers about specific articles (e.g., electronic responses for journals hosted by HighWire Press). However, few articles attract many comments, and most get none. The comments that are posted typically come from people with interest in the specific subject of the article, which means there is little overlap in the people commenting on articles in different journals. Even with comments from many peers, it remains unclear how a large set of subjective comments should be turned into a decision about scientific output.

### BASED ON ROBUST DATA

Some ventures have sought to quantify peer commentary. For example, The Faculty of 1000 maintains a large editorial board for post-publication peer review of published articles, with numerical rating being given to each rated article. Taking another approach, WebmedCentral is a journal that publishes reviewers' comments and quantitative ratings along with published articles. However, only a small fraction of published articles are evaluated by systems like these, and many of these are rated by one or two evaluators, limiting the value of this approach as a comprehensive tool for evaluating scientific contributions. It is difficult to know how many evaluations would be needed to provide a precise evaluation of an article, but the number is clearly more than the few that are currently received for most articles. Additionally, it is difficult to assess the accuracy of the comments (should one also evaluate the comments?).

It seems very unlikely that a sufficiently broad and homogeneous set of evaluations could be obtained to achieve uniformly widespread quantitative treatment of most scientists while avoiding being dominated by people who are most vocal or who have the most free time (as opposed to people with the most expertise). There is also reason for concern that peer-rating systems could be subject to manipulation (see below). For these reasons, we believe that a reliable measure of scientific output should be based on hard data rather than subjective ratings.

One could imagine specific historical instances where subjective peer commentary could have been (and probably was) quite detrimental to scientific progress. Imagine Galileo's statement that the Earth moves or Darwin's Theory of Evolution being dismissed by Twitter-like commentators.

### BASED ON DATA THAT ARE RAPIDLY UPDATED AND RETROSPECTIVE

While other sources might be useful and should not be excluded from consideration, the obvious choice for evaluation data is the citations of peer-reviewed articles. Publication of findings in peer-reviewed journals is the *sine qua non* for scientific progress, so the scientific literature is the natural place to look for a measure of scientific output. Article citations fulfill several important criteria. First, because every scientist must engage in scientific publication, a measure based on citations can be used to assess any scientist or group of scientists. Second, data on article citations are readily accessible and updated regularly, so that an index of output can be up-to-date. This may be particularly important for evaluating junior scientists, who have a short track record. Finally, publication data are available for a period that spans the lives of almost all working scientists, making it possible to track trends or monitor

career trajectories. Historical data are particularly important for validating any measure of scientific output (see below), and would be impractical to obtain historical rankings using peer ratings or other subjective approaches. Because citations provide an objective, quantifiable, and available resource, different indices can be compared (see Validation below) and incremental improvements can be made based on evaluation of their relative merits.

Citations are not without weaknesses as a basis for measuring scientific output. While more-cited articles tend to correlate with important new findings, articles can also be cited more because they contain important errors. Review articles are generally cited more than original research articles, and books or chapters are generally cited less. Although articles are now identified by type in databases, how these factors should be weighted in determining an individual's contribution would need to be carefully addressed in constructing a metric. Additionally, there will be a lag between publication and citations due to the publishing process itself and due to the time required to carry out new experiments inspired by that publication.

Citations also overlook other important components of a scientist's contribution. Scientists mentor students and postdoctorals, teach classes and give lectures, organize workshops, courses and conferences, review manuscripts and grants, generate patents, lead clinical trials, contribute methods, algorithms and data to shared repositories and reach out to the public through journalists, books, or other efforts. For this reason, subjective evaluations by well-qualified experts are likely to remain an essential component of evaluating scientific output. Some aspects of the scientific output not involving publication might be quantified and incorporated into an index of output, but some are difficult to quantify. Because it is likely that a robust index of scientific output will depend to a large extent on citation data, in the following section we restrict our discussion to citations, but without intending to exclude other data that could contribute to an index (which might be multidimensional).

We acknowledge that there are practical issues that will need to be overcome to create even the simplest metric based on citations. In particular, to perform well it will be necessary for databases to assign a unique identifier to individual authors, without which it would be impossible to evaluate anyone with names like Smith, Martin, or Nguyen. However, efforts such as these should not be a substantial obstacle and some are already underway (e.g., Author ID by PubMed or ArXiv, see [Enserink, 2009](#)).

## PRESENTED WITH DISTRIBUTIONS AND CONFIDENCE INTERVALS

An index of scientific output must be presented together with an appropriate distribution or confidence interval. Considering variation and confidence intervals is commonplace in most areas of scientific research. There is something deeply inappropriate about scientists using a measure of performance without considering its precision. A substantial component of the misuse of impact factor is the failure to consider its lack of precision (e.g., [Dimitrov et al., 2010](#)).

While the confidence intervals for an index of output for prolific senior investigators or large programs might be narrow, those for junior investigators will be appreciable because they have had less time to affect their field. Yet it is junior investigators who are

most frequently evaluated for hiring or promotion. For example, when comparing different postdoctoral candidates for a junior faculty position, it would be desirable to know the distribution of values for a given index across a large population of individuals in the same field and at the same career stage so that differences among candidates can be evaluated in the context of this distribution. Routinely providing a confidence interval with an index of performance will reveal when individuals are statistically indistinguishable and reduce the chances of misuse.

## NORMALIZATION AND FAIRNESS

### NORMALIZED BY NUMBER OF CONTRIBUTORS

When evaluating the science reported in a manuscript, the quality and significance of the work are the main consideration, and the number of authors that contributed the findings is almost irrelevant. However, the situation differs when evaluating the contributions of individuals. Clearly, if a paper has only one author, that scientist deserves more credit for the work than if that author published the same paper with 10 other authors.

Defining an appropriate way to normalize for the number of contributors is not simple. Dividing credit equally among the authors is an attractive approach, but in most cases the first author listed has contributed more to an article than other individual authors. Similarly, in some disciplines the last place in the list is usually reserved for the senior investigator, and the relative credit due to a senior investigator is not well established.

Given the importance of authorship, it would not be unreasonable to require authors to explicitly assign to each author a quantitative fractional contribution. However, divvying up author credit quantitatively would not only be extremely difficult but would also probably lead to authorship disputes on a scale well beyond those that currently occur when only the order of authors must be decided. Nevertheless, some disciplines have already taken steps in this direction, with an increasing number of journals requiring explicit statements of how each author contributed to an article.

While it seems difficult to precisely quantify how different authors contribute to a given study, if such an approach came into practice, it might not take long before disciplines established standards for assigning appropriate credit for different types of contributions. Regardless of how normalization for the number of authors is done, one likely benefit of a widely used metric normalized in this way would be the rapid elimination of honorary authorship.

### NORMALIZED BY DISCIPLINE

Scientists comprise overlapping but distinct communities that differ considerably in their size and publication habits. Publications in some disciplines include far more citations than others, either because the discipline is larger and produces more papers, or because it has a tradition of providing more comprehensive treatment of prior work (e.g., [Jemec, 2001](#); [Della Sala and Crawford, 2006](#); [Bollen et al., 2009](#); [Fersht, 2009](#)). Other factors can affect the average number of citations in an article, such as journals that restrict the number of citations that an article may include.

A simple index based on how frequently an author is cited can make investigators working in a large field that is generous with

citations appear more productive than one working in a smaller field where people save extensive references for review articles. For example, if two fields are equivalent except that articles in one field reference twice the number of articles as the other field, a simple measure based on citations could make scientists in the first field appear on average to be twice as productive as those in the second. To have maximal value, an index of output based on citations should normalize for differences in the way that citations are used in different fields (including number of people in the field, etc.). Ideally, a measure would reflect an individual's relative contribution *within* his or her field. It will be challenging to produce a method to normalize for such differences between disciplines in a rigorous and automatic way. Comprehensive treatment of this issue will require simulation and experimentation. Here, we will briefly mention potential approaches to illustrate a class of solutions.

There is a well-developed field of defining areas of science based on whether pairs of authors are cited in the same articles (author co-citation analysis; Griffith et al., 1986). More recently, these methods have been extended by automated rating of text similarity between articles (e.g., Greene et al., 2009). Methods like these might be adopted to define a community for any given scientist. With this approach, an investigator might self-define their community based on the literature that they consider most relevant, as reflected by the articles they cite in their own articles. For a robust definition that could not be easily manipulated (see below), an iterative process that used articles that cite cited articles, or articles that are cited by cited articles, would probably be needed. While it is difficult to anticipate what definition of a scientist's community might be most effective, one benefit of using objective, accessible data is that alternative definitions can be tested and refined.

Once a community of articles has been defined for an investigator, the fraction of all the citations in those articles that refer to the investigator would give a measure of the investigator's impact within that field. This might provide a much more valuable and interpretable measure than raw counts of numbers of papers or number of citations. It is conceivable that this type of analysis could also permit deeper insights. For example, it might reveal investigators who were widely cited within multiple communities, who were playing a bridging role.

### **NORMALIZED FOR CAREER STAGE**

A measure that incorporated the properties discussed so far would allow a meaningful assessment of an individual's contribution to science. It would, however, rate senior investigators as more influential than junior investigators. This is a property of many existing measures, such as total number of citations or *h*-index. For some purposes this is appropriate; investigators are frequently compared against others at a similar stage of their careers, and senior scientists generally have contributed more than junior scientists. However, for some decisions, such as judging which investigators are most productive per unit time, an adjustment for seniority is needed. Additionally, it might be revealing for a search committee to compare candidates for an Assistant Professor position with

well-known senior investigators when they entered the rank of Assistant Professor.

This type of normalization for stage of career would be difficult to achieve for several reasons. The explosive growth in the number of journals and scientists will make precise normalization difficult. Additionally, data for when individuals entered particular stages (postdoctoral, Assistant Professor, Associate Professor, Full Professor) are not widely available. A workable approximation might be possible based on the time since an author's first (or first *n*) papers were published. Because the size of different disciplines changes with time, and the rate at which articles are cited does not remain constant, these trends would need to be compensated in making comparisons over time.

A related issue is the effect of time itself on citation rates. An earlier publication has had more time to be cited (yet scientists tend to cite more recent work). In some sense, a publication from the year 2000 with 100 citations is less notable than a publication from the year 2010 with 100 citations. A simple way to address this is to compute the number of citations per year (yet we note that this involves arguable assumptions of stationarity in citation rates).

### **FOSTERING GREAT SCIENCE IMPRACTICAL TO MANIPULATE**

If a metric can be manipulated, such that it can be changed through actions that are relatively easy compared to those that it is supposed to measure, people will undoubtedly exploit that weakness. Given an index that is based on an open algorithm (and the algorithm should be open, computable and readily available), it is inevitable that scientists whose livelihoods are affected by that index will come up with ingenious ways to game the system. A good index should be impractical to game so that it encourages scientists to do good science rather than working on tactics that distort the measure.

It is for this reason that measures such as the number of times an article is downloaded cannot be used. That approach would invite the generation of an industry that would surreptitiously download specific articles many times for a fee. For the same reason, a post-publication peer-review measure that depended on evaluations from small numbers of evaluators cannot be robust when careers are at stake.

A measure that is based on the number of times an author's articles are cited should be relatively secure from gaming, assuming that the neighborhood of articles used to normalize by discipline is sufficiently large. Even a moderate-sized cartel of scientists who agreed to cite each other gratuitously would have little impact on their metrics unless their articles were so poorly cited that any manipulation would still leave them uncompetitive. Nevertheless, it seems likely that a measure based on citations should ignore self-citations and perhaps eliminate or discount citations from recent co-authors (Sala and Brooks, 2008).

One would hope that a key motivation for scientific inquiry is, as Feynman put it, "the pleasure of finding things out." Yet, any metric to evaluate scientific output establishes a certain incentive structure in the research efforts. To some extent, this is unavoidable. Ideally, the incentive structure imposed by a good metric should promote great science as opposed to incentive structures

that reward (even financially in some cases) merely publishing an article in specific journals or publishing a certain number of articles. A good metric might encourage collaborative efforts, interdisciplinary efforts, and innovative approaches. It would be important to continuously monitor and evaluate the effects of incentive structures imposed by any metric to ensure that they do not discourage important scientific efforts including interdisciplinary research, collaborations, adequate training, and mentoring of students and others.

### FOCUSED ON QUALITY OVER QUANTITY

Most existing metrics show a monotonic dependence on the number of publications. In other words, there are no “negative” citations (but perhaps there should be!). This monotonicity can promote quantity rather than quality. Consider the following example (real numbers but fictitious names). We compare authors Joe Doe and Jane Smith who work in the same research field. Both published his or her first scientific article 12 years ago and the most recent publication from each author was in 2011. Joe has published 45 manuscripts, which have been cited a total of 591 times (mean = 13.1 citations per article, median = 6 citations per article). Jane has published 14 manuscripts, which have been cited 1782 times (mean = 127.3 citations per article median = 57 citations per article). We argue that Jane’s work is more impactful in spite of the fact that her colleague has published three times more manuscripts in the same period of time. The process of publishing a manuscript has a cost in itself including the time required for the authors to do the research and report the results, the time spent by editors, reviewers, and readers to evaluate the manuscript.

In addressing this issue, care must be taken to avoid a measure that discourages scientists from reporting solid, but apparently unexciting, results. For example, penalizing the publication of possibly uninteresting manuscripts by using the average number of citations per article would be inappropriate because it would discourage the publication of any results of below-average interest. The *h*-index (and variants) constitutes an interesting attempt to emphasize quality (Hirsch, 2005). An extension of this notion would be to apply a threshold to the number of citations: publications that do not achieve a certain minimum number of citations would not count toward the overall measure of output. This threshold would have to be defined empirically and may itself be field-dependent. This may help encourage scientists to devote more time thinking about and creating excellence rather than wasting everyone’s time with publications that few consider valuable.

### VALIDATION

Given a metric, we must be able to ask how good it is. Intuitively, one could compare different metrics by selecting the one that provides a better assessment of excellence in scientific output. The argument, however, appears circular because it seems that we need to have *a priori* information about excellence to compare different possible metrics. It could be argued that the scientific community will be able to evaluate whether a metric is good or not by assessing whether it correlates well with intuitive judgments about what constitutes good science and innovative scientists. While this is

probably correct to some extent, this procedure has the potential to draw the problem back to subjective measures.

To circumvent these difficulties, one could attempt to develop quantitative criteria to evaluate the metrics themselves. One possibility is to compare each proposed quantitative metric against independent evaluations of scientific output (which may not be quantitative or readily available for every scientist). For example, Hirsch (2005) attempted to validate the *h*-index by considering Nobel laureates and showing that they typically show a relatively large *h*-index. In general, one would like to observe that the metric correlates with expert evaluations across a broad range of individuals with different degrees of productivity. While this approach seems intuitive and straightforward it suffers from bringing the problem back to subjective criteria.

An alternative may be to consider historical data. A good metric could provide *predictive* value. Imagine a set of scientists and their corresponding productivity metric values evaluated in the year 2011. We can ask how well we can predict the productivity metric values in 2011 from their corresponding values in the year 2000 or 1990. Under the assumption that the scientific productivity of a given cohort is *approximately* stationary, we expect that a useful metric would show a high degree of prediction power whereas a poor metric will not. Of course, many factors influence scientific productivity over time for a given individual and these would be only correlative and probabilistic inferences. Yet, the predictive value of a given metric could help establish a quantitative validation process.

Given the importance of evaluating scientific output, the potential for a plethora of metrics and the high-dimensional parameter landscape involved, it seems worth further examining and developing different and more sophisticated ways of validating these metrics. One could consider measures of scientific influence based on the spread of citations, the number of successful trainees, etc., and compare these to different proposed metrics. Ultimately, these are empirical questions that should be evaluated with the same rigor applied to other scientific endeavors.

### DISCUSSION

We describe above nine criteria that, we hope, might lead to a better way of evaluating scientific output. The development of an evaluation algorithm and metric that capture these properties is not intended to eliminate other forms of peer evaluation. Subjective peer review is valuable (both pre-publication and post-publication) despite its multiple pitfalls and occasional failures, and a combination of different assessments will provide more information than any one alone.

A metric that captured the properties discussed above could provide many benefits. It might encourage better publishing practices by discouraging publication of a large number of uneventful reports or reducing the emphasis on publishing in journals with high impact factors. By highlighting the scientific contributions of individuals within a field it might restore a more appropriate premium: providing important results that other scientists feel compelled to read, think about, act upon, and cite. Placing emphasis on how often other scientists cite work may have other beneficial effects. A long CV with many least-publishable papers



would quickly become visibly inferior to a shorter one with fewer but more influential papers. As mentioned above, there may be other benefits including correcting authorship practices, accurate evaluation across disciplines, and it may even help students choose a laboratory or institution for graduate studies or postdoctoral research.

In addition to evaluating the current value of a productivity metric, it may be of interest to compute the rate of change in this metric. This might help highlight individuals, laboratories, departments, or institutions that have recently excelled. Rates should also be normalized and presented alongside distributions as discussed above for the metric itself.

Although we have cast the discussion in terms of a single metric, an index of output does not need to be scalar. No single value can capture the complexities involved in scientific output. Different aspects of an investigator's contributions may require different indices. Additionally, evaluating a research group, a research center, or a department may be distinct from evaluating an individual and require somewhat different metrics (e.g., Hughes et al., 2010), but once suitable measures of output are available, productivity can be evaluated in terms of either years of effort, number of people involved, research funding, and other relevant parameters.

No calculation can take the place of a thoughtful evaluation by competent peers, and even an index that is precise and accurate can be abused. Evaluators might blindly apply an index without actually assessing papers, recommendations, and other material.

## REFERENCES

- Alberts, B., Hanson, B., and Kelner, K. L. (2008). Reviewing peer review. *Science* 321, 15.
- Amin, M., and Mabe, M. (2000). Impact factors: use and abuse. *Perspect. Publ.* 1, 1–6.
- Bollen, J., Van de Sompel, H., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE* 4, e6022. doi:10.1371/journal.pone.0006022
- Castelnuovo, G. (2008). Ditching impact factors: time for the single researcher impact factor. *BMJ* 336, 789.
- Della Sala, S., and Crawford, J. R. (2006). Impact factor as we know it handicaps neuropsychology and neuropsychologists. *Cortex* 42, 1–2.
- Dimitrov, J. D., Kaveri, S. V., and Bayry, J. (2010). Metrics: journal's impact factor skewed by a single paper. *Nature* 466, 179.
- Editors. (2006). The impact factor game. It is time to find a better way to assess the scientific literature. *PLoS Med.* 3, e291. doi:10.1371/journal.pmed.0030291
- Enserink, M. (2009). Scientific publishing. Are you ready to become a number? *Science* 323, 1662–1664.
- Fersht, A. (2009). The most influential journals: impact factor and Eigenfactor. *Proc. Natl. Acad. Sci. U.S.A.* 106, 6883–6884.
- Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA* 295, 90–93.
- Greene, D., Freyne, J., Smyth, B., and Cunningham, P. (2009). *An Analysis of Current Trends in CBR Research Using Multi-View Clustering*. Technical Report UCD-CSI-2009-03. Dublin: University College Dublin.
- Griffith, B. C., White, H. D., Drott, M. C., and Saye, J. D. (1986). Tests of methods for evaluating bibliographic databases: an analysis of the National Library of Medicine's handling of literatures in the medical behavioral sciences. *J. Am. Soc. Inf. Sci.* 37, 261–270.
- Hecht, F., Hecht, B. K., and Sandberg, A. A. (1998). The journal "impact factor": a misnamed, misleading, misused measure. *Cancer Genet. Cytogenet.* 104, 77–81.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* 102, 16569–16572.
- Hughes, M. E., Peeler, J., and Hogenesch, J. B. (2010). Network dynamics to evaluate performance of an academic institution. *Sci. Transl. Med.* 2, 53ps49.
- Jemec, G. B. (2001). Impact factor to assess academic output. *Lancet* 358, 1373.
- Mandavilli, A. (2011). Peer review: trial by Twitter. *Nature* 469, 286–287.
- Petsko, G. A. (2008). Having an impact (factor). *Genome Biol.* 9, 107.
- Refinetti, R. (2011). Publish and flourish. *Science* 331, 29.
- Sala, S. D., and Brooks, J. (2008). Multi-authors' self-citation: a further impact factor bias? *Cortex* 44, 1139–1145.
- Siegel, D., and Baveye, P. (2010). Battling the paper glut. *Science* 329, 1466.
- Simons, K. (2008). The misused impact factor. *Science* 322, 165.
- Skorka, P. (2003). How do impact factors relate to the real world? *Nature* 425, 661.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 23 June 2011; accepted: 25 October 2011; published online: 10 November 2011.

Citation: Kreiman G and Maunsell JHR (2011) Nine criteria for a measure of scientific output. *Front. Comput. Neurosci.* 5:48. doi: 10.3389/fncom.2011.00048  
Copyright © 2011 Kreiman and Maunsell. This is an open-access article subject to a non-exclusive license between the authors and Frontiers Media SA, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and other Frontiers conditions are complied with.



# Toward a new model of scientific publishing: discussion and a proposal

Dwight J. Kravitz\* and Chris I. Baker

Unit on Learning and Plasticity, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

## Edited by:

Nikolaus Kriegeskorte, Medical Research Council Cognition and Brain Sciences Unit, UK

## Reviewed by:

Thomas Boraud, Université de Bordeaux, CNRS, France  
Marc Timme, Max Planck Institute for Dynamics and Self Organization, Germany

## \*Correspondence:

Dwight J. Kravitz, Unit on Learning and Plasticity, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA.  
e-mail: kravitzd@mail.nih.gov

The current system of publishing in the biological sciences is notable for its redundancy, inconsistency, sluggishness, and opacity. These problems persist, and grow worse, because the peer review system remains focused on deciding whether or not to publish a paper in a particular journal rather than providing (1) a high-quality evaluation of scientific merit and (2) the information necessary to organize and prioritize the literature. Online access has eliminated the need for journals as distribution channels, so their primary current role is to provide authors with feedback prior to publication and a quick way for other researchers to prioritize the literature based on which journal publishes a paper. However, the feedback provided by reviewers is not focused on scientific merit but on whether to publish in a particular journal, which is generally of little use to authors and an opaque and noisy basis for prioritizing the literature. Further, each submission of a rejected manuscript requires the entire machinery of peer review to creak to life anew. This redundancy incurs delays, inconsistency, and increased burdens on authors, reviewers, and editors. Finally, reviewers have no real incentive to review well or quickly, as their performance is not tracked, let alone rewarded. One of the consistent suggestions for modifying the current peer review system is the introduction of some form of post-publication reception, and the development of a marketplace where the priority of a paper rises and falls based on its reception from the field (see other articles in this special topics). However, the information that accompanies a paper into the marketplace is as important as the marketplace's mechanics. Beyond suggestions concerning the mechanisms of reception, we propose an update to the system of publishing in which publication is guaranteed, but pre-publication peer review still occurs, giving the authors the opportunity to revise their work following a mini pre-reception from the field. This step also provides a consistent set of rankings and reviews to the marketplace, allowing for early prioritization and stabilizing its early dynamics. We further propose to improve the general quality of reviewing by providing tangible rewards to those who do it well.

**Keywords:** peer review, neuroscience, publishing

## INTRODUCTION

To begin, it is important to understand the scope and purpose of this paper. First, this paper is an attempt to describe the problems with scientific publishing as it is currently instantiated. We are both cognitive neuroscientists, and while some of the issues discussed in this paper are undoubtedly applicable to a wide array of fields they are most directly applicable to the fields of psychology and neuroscience. Second, this paper is an attempt to lay out, in a very broad way, the quantifiable and intangible costs and benefits associated with publishing so that both the functioning of the current system and the relative costs of alternatives can be evaluated. To provide some empirical basis we performed an informal survey of colleagues to obtain estimates of some of the costs. Finally, this paper includes a proposal for an alternative form of scientific publishing and post-publication review. This proposal represents our best attempt at defining an improved system that could actually be implemented given the realities of transitioning from the current system. The proposal is quite specific, but that specificity is meant

more to serve as a catalyst and basis for discussion than as a final prescription for a new form of publishing.

The paper begins with a brief discussion of the current system from an historical perspective with consideration of its modern function. Following this section is a detailed description of peer review and its tangible and intangible costs and benefits. Based on these analyses we then propose a new system for publishing empirical papers that streamlines the existing system while still serving the purposes of modern publishing. We then address the cost and benefits of this new system relative to the current system and lay out the remaining open questions.

## CURRENT SYSTEM

First, we examine the origins of the system of scientific publishing before specifying its modern form in detail. We then analyze the pragmatic, quantifiable costs of publishing based on an informal survey of 22 of our colleagues, which asked them to provide information about their experience with peer review on several of their

most recent papers (see Supplementary Material for survey), and collected information on 55 cognitive and neuroscience papers. Following this quantification of the tangible costs, we examine the intangible effects of the current system caused by the misalignment of its structure and incentives with the functions of scientific publishing.

## HISTORY AND MODERN PURPOSE

Scientific papers are published through a legacy system that was not designed to meet the needs of contemporary scientists, the demands of modern publishing, or to take advantage of current technology. The system is largely carried forward from one designed for publishers and scientists in 1665 (UK House of Commons, 2004). The most important historical constraint in shaping scientific publishing was a restriction on the available publication space. Publishing a journal, even in the recent past, was quite expensive and its likely audience quite small. Further, publishing costs are the same regardless of the quality of its content (good and bad thought costs the same to print and ship). Thus, publishers had a strong incentive to limit publication size so that the costs to readers were reasonable and to find the strongest possible content to fill that limited space. In this context, pre-publication peer review provided the publisher with a test run of the reception a paper is likely to receive from the field; providing a ranking of the likely quality of all the submitted papers. The journal then simply selected the top  $n$  papers for publication to meet its size requirement.

From the point of view of the scientists, the journals were an absolute necessity for broadly distributing their work to colleagues while still establishing ownership and precedence over a particular result (UK House of Commons, 2004). Peer review also gave scientists the same *pre-reception* it provided the journals, and with it the opportunity to revise or retract work before it was sent to the larger scientific community.

As the number of scientists grew and, concomitantly, the number of papers submitted, this system of publishing unexpectedly provided another benefit: *prioritization of the literature*. Consider the following: the price of a journal is dependent on the perceived quality of its content more than on the number of papers published. The top journal has little impetus to publish more papers as submissions increase, since by simply maintaining the number of accepted papers, the exclusivity of the journal increases and with it the perceived quality and price, with little additional expenditure (Young et al., 2008). Rejections also create a market for lower-tier journals to publish rejected papers at a reduced, but still profitable price. Scientists will naturally want to publish their work in the journal with the highest perceived quality they can, so they will submit papers to those journals first. A series of rejections and resubmissions to the next best journal will naturally lead a paper to land in the journal whose perceived quality matches that of the manuscript. Given broad agreement between scientists as to the ordering of journals by quality, and assuming that peer review is highly accurate in gauging scientific quality, the journal where a paper is published is an index to quality and thus provides its priority.

In the modern world, this prioritization and the pre-reception afforded by peer review are the primary benefits the current system of publishing provides to scientists, as the Internet has essentially

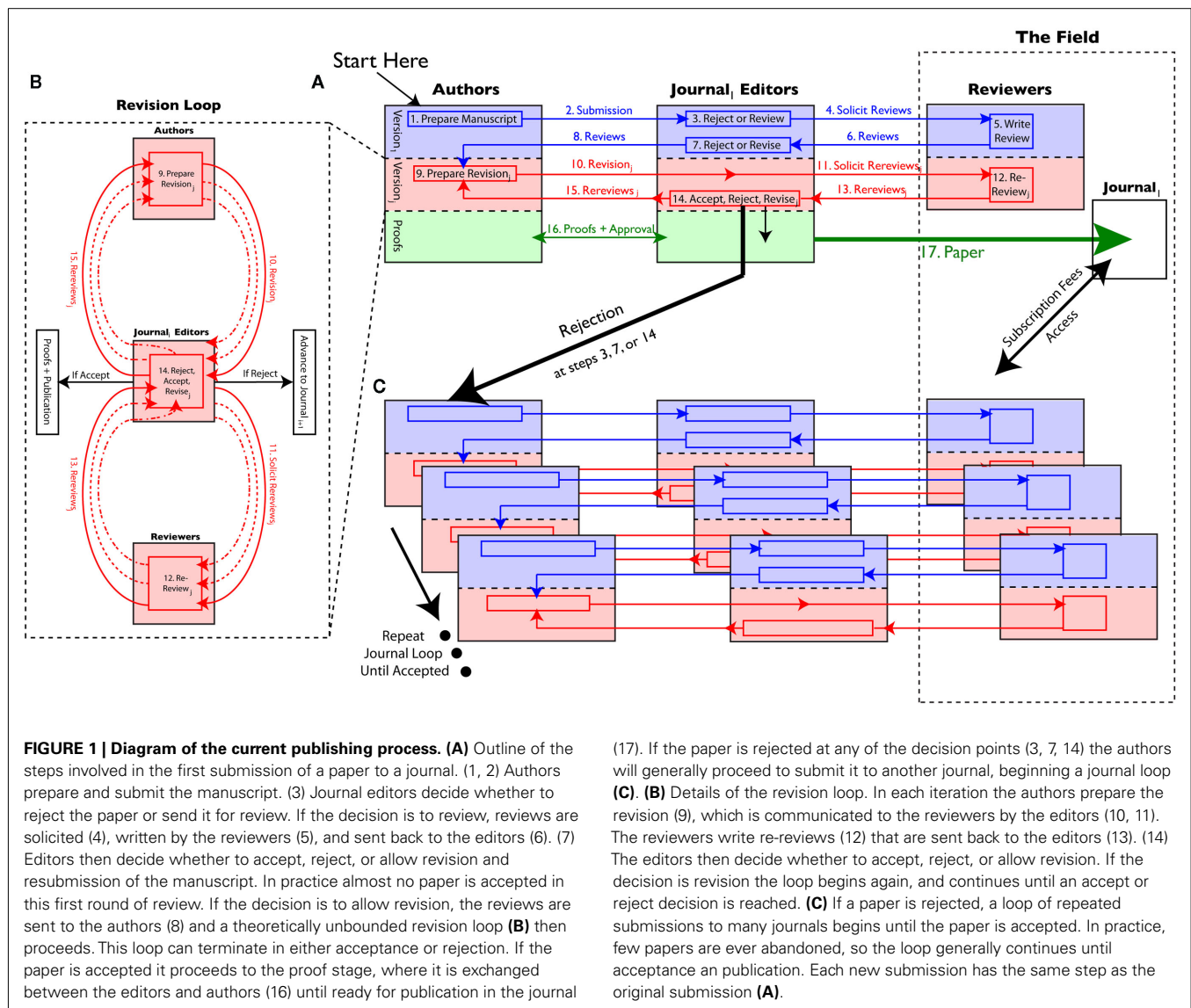
eliminated any need for journals as distribution channels. However, as the following analyses will show, the actual mechanisms of scientific publishing are poorly optimized to serve these functions.

## QUANTIFICATION OF MODERN PEER REVIEW

To effectively evaluate peer review, it is helpful to specify fully the process by which a peer-reviewed paper is currently published (Figure 1). There are three primary groups that participate in this process: *Authors* who perform research and prepare papers, *Editors* who coordinate the process of review and publication and make decisions about whether to publish or reject papers, and *Reviewers* who provide expert opinions on which the Editors base their decisions. After the initial submission by the Authors, Editors decide whether to review or reject the manuscript. If they decide to review the paper, Reviewers are solicited and, on the basis of their opinions, Editors decide whether to allow revisions to address the Reviewers' comments or to reject the paper (Figure 1A). If the decision is to allow revision, a theoretically unbounded *revision loop* begins in which the revisions pass between the three groups until the Editors ultimately reject or accept the paper (Figure 1B). In the case of a rejection the Authors generally proceed to submit the paper to a different journal, beginning a *journal loop* bounded only by the number of journals available and the dignity of the Authors (Figure 1C). When a paper is accepted, it is published in the journal and becomes available to the Field, which, for our purposes, is the set of researchers within a certain domain of research (e.g., cognitive neuroscience).

Having specified the process we can now proceed to analyze it from the point of view of its efficiency (time), cost/benefit ratio (actual expenditures of money and effort against the benefits provided), and predictability (variability in that time and effort). An ideal process maximizes the cost/benefit ratio and efficiency, while simultaneously being highly predictable. A process that is unpredictable incurs indirect costs related to the uncertainty of its function (see below).

We begin with averages representing the efficiency of the process derived from our informal survey (Table 1). There are three decision points at which Editors determine whether a paper will be rejected or continue the process at any particular journal. First, they decide whether to send papers out for review or reject them outright (26.1%; Figure 1A 3). Editors also decide whether to accept, reject, or make revisions to the manuscript following the receipt of the initial reviews (Figure 1A 7). Functionally, almost no manuscripts in our survey were accepted in the first round of review (3.6%), with most rejected (54.6%) or revised (41.8%). Once the revision loop begins, Editors repeatedly make the same accept, reject, revise decision (Figures 1A,B 14). The vast majority of papers were accepted in the same journal once the revision loop began (98.2%). Overall, however, only 33.6% of papers were published in the journal to which they were first submitted. On average, papers were submitted to 2.1 different journals (Figure 1B), underwent 2.6 revisions across all journals, and received a total of 6.3 reviews before they were published. We only collected information on papers that had been published, but it is likely that very few papers are abandoned without publication anywhere, especially given the diversity of journals now available (see also Fabiato, 1994; Suls and Martin, 2009).



Beyond these raw numbers our survey also provided us with estimates of the amount of time taken in various steps of the process. Here, what is striking is less the average amount of time, which is quite long, but more its unpredictability. In total, each paper was under review for an average of 122 days but with a minimum of 31 days and a maximum of 321. The average time between the first submission and acceptance, including time for revisions by the authors was 221 days (range: 31–533). This uncertainty in time makes it difficult to schedule and predict the outcome of large research projects. For example, it is difficult to be certain whether a novel result will be published before a competitor's even it were submitted first, or to know when follow up studies can be published. It also makes it difficult for junior researchers to plan their careers, as job applications and tenure are dependent on having published papers.

We also asked for the amount of time taken to prepare submissions and reviews, allowing us to estimate the actual work and expenditure consumed in the process. Leaving aside the initial

preparation of the paper (Figure 1A 1) we begin with the preparation of reviews (Figure 1A 5). Each paper received, on average, 6.3 reviews and, each review takes, on average, 6 h to prepare (based on an informal survey of post-docs in our lab). At the average salary for a NIH post-doc (\$47,130 for approximately 2000 yearly hours<sup>1</sup>), this roughly translates to a cost of \$140 per review and \$840 per paper. Importantly, these reviews will never be seen outside of the review process, so their only utility is in refining published manuscripts. Next we consider the preparation of revisions and submissions to different journals. In our survey, Authors estimated that they spent, on average, 68 h on all the revisions and resubmissions, roughly translating to a cost of \$1600 per paper prior to acceptance. While these estimates of time spent may not be highly precise, they do provide a rough basis for estimating the

<sup>1</sup> [http://www.glassdoor.com/Salary/NIH-Postdoctoral-Fellow-Salaries-E11709\\_D\\_KO4,23.htm](http://www.glassdoor.com/Salary/NIH-Postdoctoral-Fellow-Salaries-E11709_D_KO4,23.htm)



**Table 1 | Summary of survey statistics.**

1. % First submissions rejected without review (Figure 1A 3)	26.1
2. % First submissions rejected/revised/accepted after review (Figure 1A 7)	54.6/41.8/3.6
3. % Papers rejected/accepted in revision loop (Figure 1A 14)	1.8/98.2
4. % Papers published in the first journal	33.6
5. Average total journals (Figure 1C)	2.1 (1–6)
6. Average total revisions (Figure 1B)	2.6 (1–6)
7. Average total reviews	6.3 (2–15)
8. Average total time under review (days)	122 (31–321)
9. Average estimated total time to prepare revisions (hours)	68 (5–300)
10. Average time from first submission to publication (days)	221 (31–533)

Each of the measures is based on a survey of 55 papers from 22 individual researchers. 1. Gives the percentage of first submissions to any journal that were rejected without review by the Editor. 2. Gives the percent of reviewed first submissions that were given a decision of reject, revise, or accept. 3. Gives the percent of papers that were accepted or rejected at a journal once they were given a revise decision. 4. Gives the percent of papers that ended up published in the first journal to which they were submitted. 5. Gives the average total number of journals to which the papers were submitted. The number in parentheses gives the range. 6. Gives the average number of revisions a paper underwent across all journals excluding first submissions. 7. Gives the average total number of reviews that were done for each paper. 8. Gives the average total amount of time in days the paper was under review across all submissions. 9. Gives the average estimated time in hours to prepare all the revisions of a paper. 10. Gives the average time in days between the first submission to an accept decision.

total cost. Finally, (based on the last few publications from our lab) the average direct cost of publishing a paper in terms of publication fees (e.g., color figure costs) was \$1930. Beyond the costs of actually performing the research and preparing the first draft of the manuscript, it costs the field of neuroscience, and ultimately the funding agencies, approximately \$4370 per paper and \$9.2 million over the approximately 2100 neuroscience papers published last year. This excludes the substantial expense of the journal subscriptions required to actually read the research the field produces and the unquantifiable cost of the publishing lag (221 days) and the uncertainty incurred by that delay.

### INTANGIBLE COSTS AND BENEFITS

Given these costs, we now turn to evaluating the functionality provided by the current system to the field, which ultimately funds its every component. Beyond the ineffectiveness of the current system in providing pre-reception and a prioritization of the literature, we also highlight the costs caused by the misalignment of incentives and the adversarial relationship between the Reviewers and Authors caused in the current system.

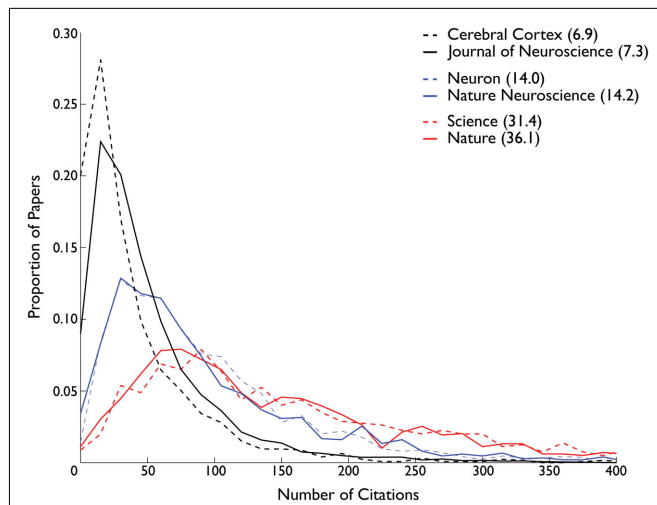
The current system serves the purposes of the journals, providing them with a *pre-reception* that allows them to prioritize papers for publication. However, this pre-reception is ill-suited to needs of scientists as it is optimized to help the journals decide whether or not to publish and not for providing feedback about scientific merit. Further, because the sample of Reviewers is so small relative to the size of the Field, and their identities generally unknown, it

is very hard for Authors to know how general the Reviewers' opinions will be in all but the most extreme cases. Reviewers may also be implicitly biased in their reviews by their feelings about particular Authors. One study (Peters and Ceci, 1982) resubmitted 12 articles already published in high-tier journals with different authors names and institutions. First, only three of the papers were detected as already published, and at a time when the number of published papers was much lower than it is today. Second, eight of the nine remaining papers were rejected, none for novelty, but generally for "serious methodological flaws." This result might suggest a systematic bias by Reviewers or that peer review itself is unreliable. In either case, this form of pre-reception is clearly not optimal for Authors.

The *prioritization of the literature* afforded by this system is also quite poor. From the point of view of the Authors, the system is so stochastic and redundant as to be an active hindrance to the progress of research. The redundancy also increases the burden on Reviewers, who are essentially uncompensated, as the same paper requires a multitude of reviews through the revision and journal loops. From the point of view of an individual researcher in the field, there is no guarantee that the criteria of a journal or those of the Reviewers match their own, especially in the case of the highest tier journals in which novelty plays a large part in the decision to publish. Not only is novelty inherently subjective, the question is being asked of specialists who are unlikely to have a good intuition of novelty or general interest in the larger scientific community. Further, the general novelty of a result may have little to do with its actual importance to the research program of any particular researcher. Thus the prioritization of the literature provided by this process is, at best, noisy, opaque, and very expensive.

To quantitatively evaluate the performance of the journals in prioritizing the literature, we extracted from SCOPUS the current number of citations for all the neuroscience papers published between 2000 and 2007 in six major journals. The journals were chosen from three distinct tiers based on impact factor, the dominant measure of journal quality used in the field. If the journal is a good marker of a paper's quality and eventual impact on the field, then the eventual citation count of that paper should be predictable from the journal where it was published. Viewed retrospectively, this should lead to largely non-overlapping distributions of citation counts between journals in different tiers. It should be noted that this measure is somewhat confounded by the fact that high-tier journals are both more visible and more likely to attract submissions than lower-tier journals. However, both of these confounds should act to increase distinctions between the tiers. Our evaluation reveals that far from a perfect filter, the distribution of citations largely overlaps across all six journals (Figure 2). We then asked whether the citation count of a paper could predict the tier at which it was published and found that between adjacent tiers this could only be achieved at 66% accuracy and between the top and third tier at 79%<sup>2</sup>. Thus, even given

<sup>2</sup>This calculation was achieved by drawing every possible boundary in citation count and assessing the proportion of the distribution for each journal that fell on either side of the boundary. Subtracting the proportion of the each journal that fell on the same side of the boundary from one another provides the percent correct for a particular boundary. The percent correct from the best boundary is reported.



**FIGURE 2 | Prioritization of the literature by the current system.**

Histogram of the distribution of the current number of citations for every neuroscience paper published between 2000 and 2007 for six major journals (15 citation width bins). The x-axis is cutoff at 400 citations only for display purposes. There were a small proportion of papers that had more citations, and these papers were included in all analyses. There are three rough tiers of journals, based on their 2010 impact factors (to the right of the journal names in the legend). Note the large amount of overlap between the distributions; indicating the journal where a paper is published is not strongly predictive of the eventual number of citations it will acquire.

the self-reinforcing confounds, the journals tiers are far from a perfect method of prioritizing the literature.

The current system is also notable for the misalignment of incentives for both Authors and Reviewers relative to progress in science. Scientific progress is supposed to be largely incremental, with each new result fully contextualized with the extant literature and fully explored with many different analyses and manipulations. Replications, with even the tiniest additional manipulations, are critical to refining our understanding of the implications of any result. Yet, with the focus on the worthiness for publication, especially novelty, rather than on scientific merit, Reviewers look on strong links with previous literature as a weakness rather than strength. Authors are incentivized to highlight the novelty of a result, often to the detriment of linking it with the previous literature or overarching theoretical frameworks. Worse still, the novelty constraint disincentivizes even performing incremental research or replications, as they cost just as much as running novel studies and will likely not be published in high-tier journals.

The current system also creates an adversarial relationship between Reviewers and Authors. Asking Reviewers to make judgments about publication worthiness reduces criticism to a dichotomy: Accept or Reject. Most of the comments in reviews reduce to this boolean, so Authors are incentivized not to argue or discuss points but simply to do enough to get a paper past the Reviewers. Reviewers are essentially uncompensated and completely anonymous, so there is no incentive to produce timely, let alone detailed constructive reviews. To Authors, a review often reduces to a list of tasks rather than as a scientific critique or

discussion that refines a paper. In practice, most reviews are rejections or lists of control experiments that are often not central to the theoretical point being addressed which bloat papers rather than refining them. To be clear, these problems occur even with the most conscientious Reviewers, which most researchers try to be, simply because of the nature of the current system of publishing. With no reward for or training in good reviewing and counter-productive incentives, it is unsurprising that peer review is ineffective at producing either a high-quality pre-reception or a prioritization of the literature.

## PROPOSED SYSTEM OF PEER REVIEW

Luckily, these deficiencies are structural and do not arise because of evil Authors, Reviewers, or Editors. Rather, they are largely a symptom of the legacy system of scientific publishing, which grew from a constraint on the amount of physical space available in journals. The advent of the Internet eliminates the need for physical copies of journals and with it any real space restrictions. In fact, none of the researchers in our lab had read a physical copy of a journal in the past year that was not sent to them for free. Without the space constraint there is no need to deny publication for any but the most egregiously unscientific of papers. In fact, we argue that simply guaranteeing publication for any scientifically valid empirical manuscript attenuates all of the intangible and quantifiable costs described above. Functionally, publication is already guaranteed, it is simply accomplished through a very inefficient system. 98.2% of all papers that enter the revision loop are published at that same journal and few papers are abandoned over the course of the journal loop.

Guaranteeing publication would dramatically simplify the process of peer review, align the incentives of Authors and Reviewers with scientific progress, and reduce costs in time, money, effort, and uncertainty. In our detailed description of our proposed system (see below), we will even show that guaranteed publication does not sacrifice, and in fact, improves both *pre-reception* and the *prioritization of the literature*. We begin with a specification of the mechanisms and costs of the proposed system, followed by a discussion of the intangible costs and benefits. A high level summary can be found in **Table 2**.

## PROPOSED SYSTEM OF PEER REVIEW AND QUANTIFICATION

Guaranteeing publication would eliminate the redundancy of the revision and journal loops, improving every quantifiable aspect of peer review. Under the proposed system (**Figure 3**) all papers are reviewed. The purpose of the Editors is twofold. First, they coordinate the entire review process. Second, they maintain the anonymity of both the Reviewers and Authors, so that all reviewing is double-blind (see Peters and Ceci, 1984 for a discussion). Editors pick a set of three anonymous reviewers based on their expertise and availability (**Figure 3**). Once the reviews are prepared, they are passed automatically to the Authors, without the need for any editorial decisions (**Figure 3**). The purpose of these reviews is not to decide whether the paper should be published, but to give the Authors feedback on the scientific quality of the research and the Reviewer's understanding of its context and importance in the field. This scientific pre-reception affords the Authors the opportunity to significantly revise or retract their work if they

**Table 2 | This table contains a rough summary of the key differences between the current and proposed systems of peer review.**

CURRENT SYSTEM	
<b>Limits publication based on a non-existent space constraint</b>	
Pre-reception	Encourages reviews focused on publication rather than scientific merit Untracked and unrewarded reviewers have no incentive to review well
Prioritization of the literature	Static and based on which journal publishes a paper Limits competition between journals for papers Creates long publication lags
Other problems	Disincentivizes incremental research Introduces uncertainty in publishing time Provides no medium for rapid and ongoing discussion of paper
PROPOSED SYSTEM	
<b>Guarantees publication of valid research</b>	
Pre-reception	Reviews focus only on scientific merit Tracks and rewards reviewers
Prioritization of the literature	Ongoing and flexible evaluation of papers even after publication Editors directly compete for papers Fixed and short publication lag
Other costs	Drastically reduces uncertainty in publishing time Reduces the money and effort expended by the field Post-publication system provides for public discussion and clarifications

choose. If both the Authors and Reviewers agree multiple rounds of review are possible (e.g., Frontiers system). However, in most cases, the Authors will instead respond to the reviews once and make some revisions to their manuscript (Figure 3 6 + 7). Having communicated that revision to the Editors, publication is now guaranteed with no further rounds of revision or review. The elimination of the revision and journal loops significantly reduces the inefficiency and speed of publication but a method is still required for prioritizing the literature.

To this end, we propose combining post-publication review (see below) with an Editorial Board, whose function is to provide initial seeds that will be the basis of the early prioritization of papers as they are published. The Editorial Board essentially acts as a rating service, fashioning a coherent summary and set of ratings from the raw initial reviews and responses that the field can use to initially prioritize a paper. The Board will be comprised of a small set of leaders in the field, chosen, at least initially (see below), by vote amongst the field. Editors will send the paper, reviews, and responses (all anonymous) to a primary member of the Editorial Board, who will be responsible for providing the initial ratings and summary, including their own impressions of the implications of the paper in context with the literature (see also Faculty of 1000 for a related system; Figure 3 9). Once these seeds are complete and the proofs receive final approval (Figure 3 11), the paper is immediately published.

The proposed system will immediately reduce the burden in time, money, and effort on the entire field. Given a single round of review, the number of reviews is reduced from a current average of 6.3 to 3 saving the field 18.2 h of reviewing and \$430 per paper on average (52%). There is only a single optional round of revision, saving Authors an average of 42 h and \$990 per paper on average (62%) according to our survey. Even assuming that the publication and submission fees remain constant to pay for the implementation of the new system (and color figure fees would certainly be eliminated), a total savings of \$1420 would be achieved for each paper (32%). That translates to an annual savings of three million dollars for the field, not including the benefits of a reduction in publication lag and the decrease in uncertainty.

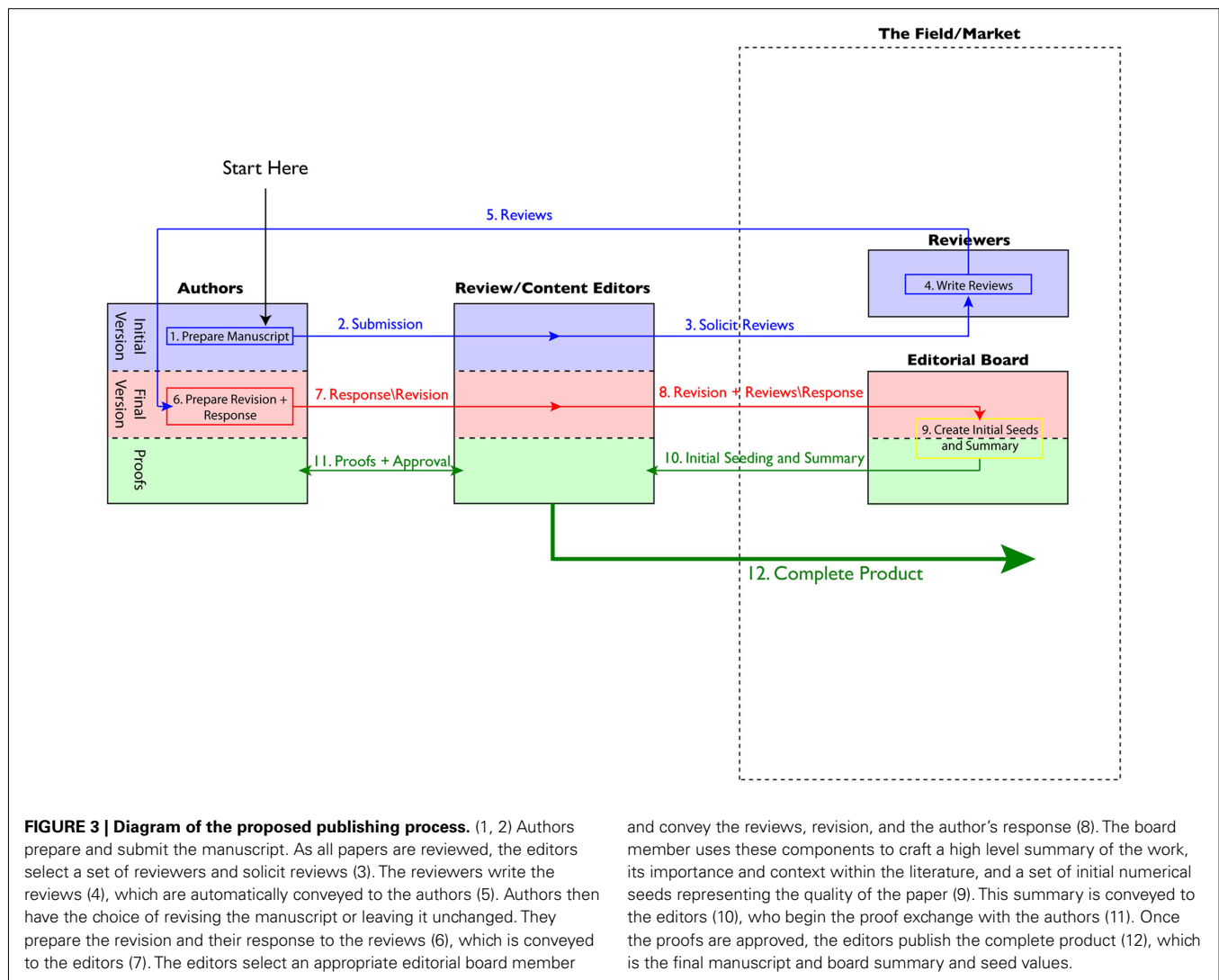
### INTANGIBLE COSTS AND BENEFITS

The proposed system of peer review streamlines the existing system, benefiting Authors without fundamentally changing their role. Authors continue to perform research and write papers, but a greater proportion of their time can now be devoted to actually doing those things. They are also the beneficiaries of an improved, more scientific *pre-reception* and a reduced cost and lag for publishing papers. The reduced variability in time reduces uncertainty, helping junior scientists plan their careers more effectively, and helping senior researchers plan large-scale research projects.

The role of Reviewers is altered from assessing publication worthiness to providing a critique of the paper's scientific merit. This should reduce the adversarial relationship between Authors and Reviewers, and foster more constructive criticism. When this system is combined with an appropriate system of post-publication review, it may also provide Reviewers with the opportunity to be directly rewarded for producing high-quality punctual reviews (see Compensation for Reviewers section below).

For Editors, the change will be fundamental. Currently, Editors are the gatekeepers to publication in a particular journal. Their purpose is to serve the interests of the journal as a business and not the interests of Authors. There is also no real competition between Editors, as the entire system rests on a relatively well-established hierarchy of journals to provide the prioritization of the literature. In the proposed system, journals do not truly exist as distinct entities for the purposes of peer review (though they may play a role in post-publication as discussed below in the Financing section). Instead, Editors must function in a way somewhat analogous to an investment bank, shepherding a paper into the market in its best possible form. Editors can compete with each other based on the price and quality of the services they provide. For example, Editors can both coordinate the pre-publication review process, and more or less extensively edit the manuscript and figures, provide digestible press releases for high-profile papers, and promote the manuscript within the community. The Nature publishing group has started offering a variant of this service already, by offering to edit manuscripts they will not necessarily publish<sup>3</sup>. The proposed system aligns Editor's incentives with the desire of the Authors to publish the best possible paper in a certain time frame with a reasonable cost.

<sup>3</sup><http://languageediting.nature.com/>



and convey the reviews, revision, and the author's response (8). The board member uses these components to craft a high level summary of the work, its importance and context within the literature, and a set of initial numerical seeds representing the quality of the paper (9). This summary is conveyed to the editors (10), who begin the proof exchange with the authors (11). Once the proofs are approved, the editors publish the complete product (12), which is the final manuscript and board summary and seed values.

### REASONS FOR DOUBLE-BLIND PRE-PUBLICATION REVIEW

Unlike many other proposals we propose maintaining some pre-publication peer review. While eliminating this step would further simplify and streamline publishing we believe it to be critical for three reasons. First, review by experts in the field prior to publication is critical for providing the Authors with an effective pre-reception that can be the basis for revising or retracting papers before they become widely available. Second, the reviews, once synthesized by the editorial board, can also serve as an early input into the post-publication market, stabilizing initial reception. Third, it also guarantees that every paper will receive an initial set of reviews, eliminating the concern that a paper that is never commented on post-publication is essentially invisible to any prioritization (see also below).

We further argue that this pre-publication review should be double-blind, with the identities of both the Authors and Reviewers unknown to the other. The anonymity of the Reviewers is critical to obtaining unadulterated reviews, particularly when more junior scientists are reviewing the work of senior faculty (e.g., Wright, 1994). In cases of completely open peer review, reviews

become more positive and acceptances increase (Van Rooyen et al., 1999), but so does hesitancy to review in the first case. It is unclear whether the increased positivity reflects genuine enthusiasm or merely the desire to avoid conflict. The anonymity of the Authors reduces the possibility of Reviewer bias either for or against particular authors or institutions (see Peters and Ceci, 1982 for an example). While the identity of the Authors might be guessed by the Reviewers, any ambiguity should act to reduce this bias.

### REASONS FOR INCLUDING AN EDITORIAL BOARD

Beyond streamlining the existing system of peer review we propose the addition of an Editorial Board, responsible for preparing a summary based on the initial reviews and a set of initial ratings that accompany a paper as it is published into the market. The inclusion of this group adds steps and time to the process of publication and also creates a new burden on the field. Nonetheless, the benefits of the Editorial Board outweigh these costs.

Current systems that depend on post-publication review are plagued by an uneven initial reception. Complete post-publication review puts an enormous burden on the field to conscientiously

search the literature and offer commentary without any compensation whatsoever (Lipworth et al., 2011). The only researchers likely to offer comments are those deeply invested in a particular result, and there is little point in offering positive commentary on a paper. In current open review systems, some papers are commented on extensively, while others never receive a single comment (e.g., Nature open peer review debate<sup>4</sup>). The latter case provides neither the field nor the Authors any sort of feedback on the quality of the research, nor any prioritization of the literature. The Editorial Board provides ratings and a summary that can provide an early prioritization of papers and guarantee that every paper is read and contextualized with the extant literature.

The Editorial Board offers significant advantages over publishing the raw initial reviews with the paper. First, many of the initial issues will be fully addressed by the response and will add nothing to the early reception of the paper. Second, publishing the reviews would tend to recreate the adversarial relationship between the Authors and Reviewers, as the Reviewers would be implicitly accepting the Authors' response without the opportunity to argue their points or to revise their review. The inclusion of an impartial third party to provide the final word on whether issues have been addressed or remain outstanding, gives both the Reviewers and Authors some distance from their reviews and responses. Finally, the Editorial Board can also evaluate the quality and timeliness of reviews, perhaps providing a metric on the basis of which Reviewers can be rewarded (see Compensation for Reviewers section below).

## PROPOSED POST-PUBLICATION SYSTEM

There are four primary functions that the structure of a paper must serve if it is to be considered effective. (1) It must convey the content of the research in such a way that it can be understood and replicated. The existing structure of published papers is well-established and entirely sufficient to accomplish this goal. (2) It must provide a way to contact Authors for clarifications. (3) The structure must provide an easy method for indexing the paper in relation to the issues it addresses and the rest of the literature. Currently, this indexing is accomplished through the combination of keyword searches and citation linkages. (4) It must have a set of statistics and comments associated with it that allow its reception by the field to be tracked for the purposes of evaluating individuals for funding and promotions and prioritizing it within the literature. Some journals and search engines have already begun to track download count and number of citations. While the current structure has been adapted to serve these functions, it is far from optimal, and online access allows us the opportunity to design a new structure with superior functionality.

## STRUCTURE OF PAPERS POST-PUBLICATION

Under the proposed system, a published paper will consist of the following components. First, the article itself (**Figure 4A**, green box), which has essentially the same structure as papers currently have with the addition of the summary and initial ratings from

the Editorial Board. The original article will be the only component that is immutable – once published it will never change. This component provides a consistent way for Authors to claim work as their own, and the familiar format of the article is ideally suited to serving the first function. The only major change in this structure will be that the format will be consistent. We will, as a Field, decide on a common list of components (e.g., abstract, introduction, etc.) and stick to it, rather than reformatting manuscripts for each journal.

Second, the paper will also be associated with a forum (**Figure 4A**, purple circle) within which members of the field can ask methodological and theoretical questions as well as offer up their own detailed reviews. Upon publication, the original Reviewers will be invited to anonymously post their reviews (with any modifications) if they choose, but all other contributions will be open and directly associated with particular researchers. Authors are free to respond to any post in the forum, adding comments, or additional data as appropriate, as are members of the field in general. The forum provides a way for Authors to publicly refine their work and theories as the process of reception unfolds without needing to publish new papers on minor incremental or clarifying points. The forum also provides a way for the field to reach a consensus on the implications and limitations of any result. Critically, the forums provide a record of these discussions, again providing Authors the ability to claim at least informal ownership of particular ideas outside of the context of published papers.

The paper will also contain a set of continuously updated statistics (**Figure 4A**, yellow circle) which track the reception the paper is receiving from the field. These statistics are essentially numeric data that provide an easy way of prioritizing the paper by tracking things like citation and download counts. They also include ratings provided by members of the field after publication.

Finally, all of these components along with some additional information comprise a literature valence (**Figure 4A**, large blue circle), which can be used to both prioritize the paper and place it in context with the literature. The additional information includes other work that has cited the paper since publication, the IDs of the Authors, Reviewers, Editors, and Editorial Board member, keywords, and additional related literature suggested by them or any other members of the field.

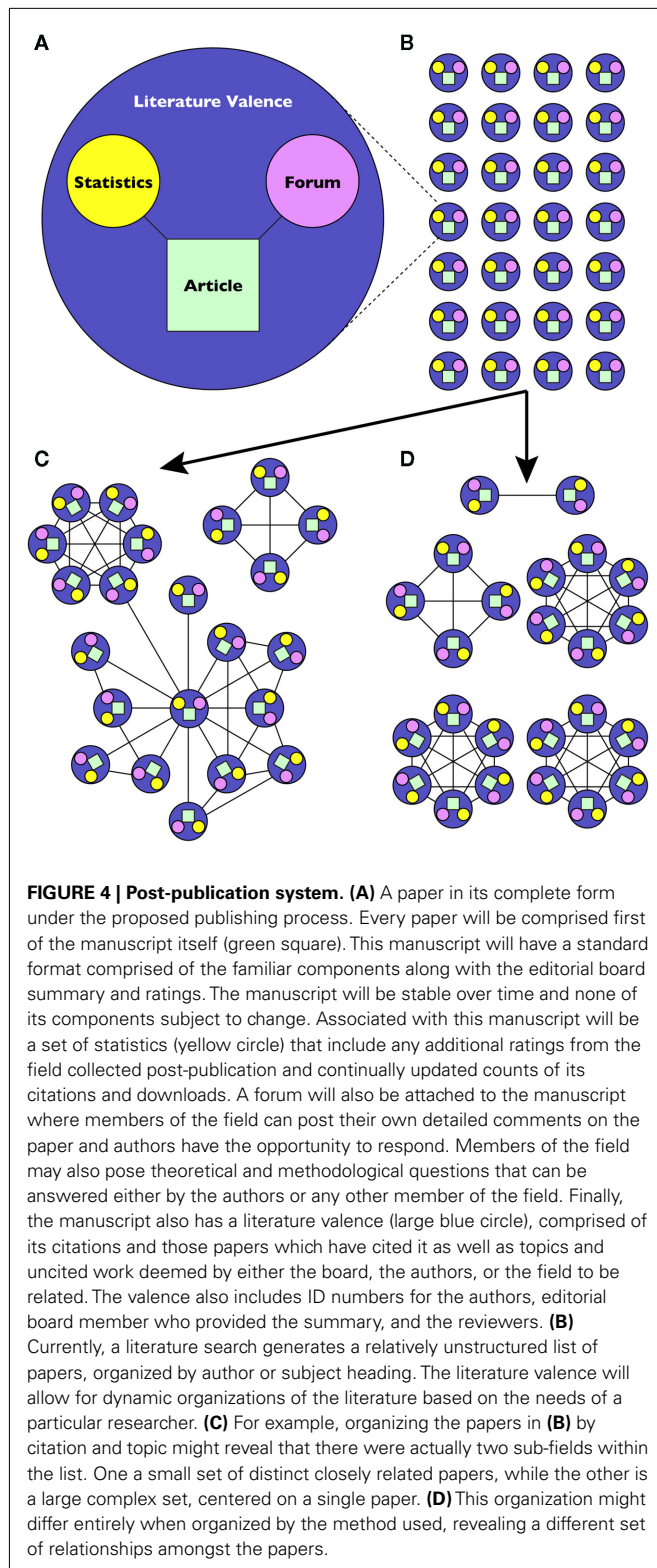
The structure described above is relatively similar to the backbone of social networking websites like Facebook. The problems being addressed by the two systems is similar in that both create a virtual anchor for an actual person or paper, to which content can be continuously added and indexed online without altering the fundamental link between the anchor and the actual content. In fact, beyond the statistics tracking, most of the functionality described above can be achieved simply by making a Facebook page for a paper. This similarity is a strength of the proposed system as it dramatically simplifies and cheapens the implementation of the proposed system (see Finance section below).

## SEARCHING AND ORGANIZING THE LITERATURE

When the information in the literature valence is married to the appropriate algorithms it can yield a very powerful and flexible

<sup>4</sup><http://www.nature.com/nature/peerreview/debate/index.html>





way of organizing the literature. For example, literature searches currently yield a list of papers associated with a particular keyword or author, generally ordered by date (Figure 4B). While this organization is useful as a first pass, an additional algorithm

which takes into account the citations might reveal a much more informative structure: in this example, two distinct subgroups of papers with one subgroup being centered on a single seminal paper (Figure 4C). Alternatively, an organization based on the methods used (e.g., fMRI) might show an entirely different grouping, with many different methods being used to address the same topic (Figure 4D). The proposed system would also allow searches and organizations based on who reviewed the paper, which editorial board member wrote the summary, or the post-publication ratings of a particular individual researcher. The point is not the particular organization but to build a structure flexible enough to support a wide range of organizations tailor-made to the needs of individual researchers.

## OPEN QUESTIONS

### FINANCING THE SYSTEM AND TRANSITION FROM THE CURRENT SYSTEM

In the preceding sections we proposed a new system of publishing that does not completely demolish the existing system but streamlines it and optimizes it to leverage the currently available technology. This approach is critical, as it leads to a new system that can be easily and cheaply transitioned to from the current system. In this section we review the major components of the proposed system that will require expenditures of money and effort to implement and maintain.

First, there is the coordination of the review process. Currently, this function is served by journals that are financed by a combination of subscription, submission, advertising, and publication fees. In the proposed system, the editorial process is decoupled from publication, all published papers are freely available, and physical copies of journals are no longer produced. This reduces the source of revenue for the editorial process essentially to submission fees provided by the Authors. There are, however, several factors that will attenuate these costs. (1) Publication is guaranteed, so payment of the fee will definitely lead to a publication. (2) Editors will now have to directly compete with one another on the basis of price and quality of service (i.e., speed, copy editing, publicity for high-profile results). This competition should lead to a wide range of Editor pricing and services and should reduce fees overall. (3) It is likely advisable to have a single electronic backbone that is used for the coordination of Reviewers and the Editorial Board. This system could track the number of papers currently assigned to individuals, making the assignment of new papers more efficient. It would also eliminate redundant implementations of similar systems by different Editors, and provide a common set of anonymous IDs for Reviewers across all submissions. All of these factors should increase efficiency and reduce the overall price. The implementation and maintenance of such a system is quite simple and could be easily paid for from a general funding source (e.g., NIH) or by a proportion of the submission fees. The transition to this system of pre-publication review will probably need to be done as a field, as the proposed system would be hard-pressed to compete with the more prestigious journals that already exist. The other alternative is to create such a system and wait for its increased efficiency to render the other modes of publication obsolete over a likely period of many years.

Second, the backbone of the post-publication market must be implemented and maintained. Again, it is likely advisable that a single system serves the entire field, to maintain consistency, reduce redundancy, and provide a common access point for the literature. A single system could also be used to track all users and to restrict access to accredited institutions and individuals or to ban users who abuse the system if needs be. Since the basic structure of the proposed post-publication market is similar to existing social networking sites, the minor extensions required would not be overly costly to implement or maintain for these companies. Revenue could be generated by again taking a proportion of the submission fees. It could also be generated through targeted advertisements. The topic headings of papers provide an excellent index into the scientific apparatus likely needed by researchers reading that paper. Whereas currently most advertisements for these products are scattershot, pushed through journals or emails, associating the ads with particular papers might be more effective. Another advantage of the proposed post-publication market is that it can be implemented independent of the proposed system of pre-publication review. Even existing papers can be adapted into the proposed marketplace and their reception tracked, easing the transition to the proposed system.

Finally, the front-end service by which the literature can be searched, organized, prioritized, and presented to researchers will need to be funded. Currently, there are a number of search engines (e.g., Pubmed), financed by the major funding agencies that could be adapted to serve these functions. However, this is also a potential market for the existing journals, which could provide several distinct services to scientists. (1) Journals can produce their own proprietary prioritizations of the literature. In the proposed system any prioritization essentially reduces to some, likely linear, formula representing a combination of all the available factors. That equation can be proprietary and journals can offer their own prioritizations to researchers for a fee. In fact, some journals have already begun to offer something similar to this function, by providing field-wide research highlights with every published issue. This can lead to the strange experience of being rejected by a journal and then having the same paper highlighted within it later. (2) Similarly, journals can offer new algorithms for organizing the literature; perhaps even offering a direct service to researchers. (3) Journals might also be the logical outlet for review articles, which would be trivial to publish under the proposed system. If review articles were limited to invited pieces in particular journals, they could be published under a different system more directly suited for them. Journals could also charge for access to these articles just as they charge for empirical pieces currently.

### COMPENSATION FOR REVIEWERS

Our proposed system reduces the reviewing burden on the field and better aligns the incentives, but we recognize that our proposed system is still dependent on the efficiency and quality of the reviews. Unless reviewing is directly rewarded, it will always be at the bottom of the stack for any researcher. Further, we, as a Field, need to acknowledge the importance of reviewing as part of doing good science and reward researchers for doing it well. In the current system, good reviewing is not even defined, let alone tracked,

and it is the backbone of all publishing. Finding a way to track and reward good reviewing might also reveal a heretofore-unknown group of researchers who are gifted in it and might teach the rest of us how to do it effectively.

Our proposed system provides mechanisms that allow reviewing to be tracked and rewarded. The raw initial reviews are provided to the Editorial Board, whose members could be asked to rate the usefulness and insightfulness of those reviews. Assuming that the identity of the Reviewers is kept anonymous, this could provide a relatively unbiased estimate of the quality of the reviews, similar to a system already in place at some journals (e.g., PLoS ONE). Upon publication, the Reviewers could also be asked to provide final ratings that could be regressed against the actual reception of the paper and final reviews that could be rated by the field.

Having tracked the quality of individual Reviewers, the question is how best to reward them. First, statistics representing the quality of a Reviewer could be cited in job applications and tenure reviews. Second, high-quality reviewing could qualify a Reviewer for membership in the Editorial Board (see below). Finally, Reviewers could also be paid a proportion of the submission fees commensurate with the quality of their reviewing for each paper they review. These fees would not have to be paid to Reviewers directly, instead they could be added to existing grants in the Reviewers lab, or could simply defray submission costs for the Reviewer's own papers.

### MECHANISMS OF THE EDITORIAL BOARD

Under the proposed system the Editorial Board has a very important responsibility to provide the initial summary and ratings that accompany a paper into the marketplace. Beyond this responsibility, members of the Editorial Board also have the burden of producing these summaries and ratings for every published paper. As such, the size of Board, its membership, and compensation for serving on it must be carefully considered.

The size of the Board is the least complicated of the issues. All that is required is to ascertain the average amount of time it takes to produce a summary and a set of initial rankings for each paper. Assuming that this process is comparable to reviewing a paper (6 h), an Editorial Board member could reasonably handle two papers a week. Dividing the number of papers submitted in a week ( $\sim 40$ )<sup>5</sup> by this number would yield a rough estimate of the necessary size of the Editorial Board ( $\sim 20$ ). This number could then be adjusted after the system begins operation. Alternates could also be specified who could contribute during times with very high numbers of submissions.

The membership of the Editorial Board is a more complex issue. Initially members should probably be elected to some set terms by the members of the field. Once those terms end or members resign, they can also be replaced by a voting procedure. Some positions might also be filled by the best Reviewers in the field (see Compensation for Reviewers section above), providing another reward for good reviewing.

<sup>5</sup>This number was calculated by dividing the number of papers with the topic "neuroscience" published in 2010 (2100) by 52 weeks.

Finally, serving on the Editorial Board incurs a significant cost in both time and effort and its members will need to be compensated. On the one hand, serving on the Editorial Board will be very prestigious and the position provides the opportunity to help shape the direction of the field, so in some sense serving is its own reward. On the other hand, members could also receive some direct compensation, likely in the form of some guaranteed funding for their labs. This would remove members from the grant treadmill, freeing them to more fully immerse themselves in the literature. Further, it would reduce the burden on grant reviewers, who would no longer have to review grants that are very likely to be funded (particularly if membership in the Editorial Board is determined by voting).

## CONCLUSION

Ultimately, the process of reforming the current system of publishing will be long, arduous, and fraught with uncertainty. The purpose of this manuscript is not to propose a final solution; by no means is the proposed system perfect. Instead we sought to highlight the problems in the current system, the functions that should guide the new system, and the necessity of reforming the system (see **Table 2**). It is to this final point that we now turn in some additional detail. Above, we have argued, in some depth, that the current system is needlessly redundant, expensive, and ill-suited to meeting the needs of the field, specifically a scientific *pre-reception* for Authors, and a *prioritization of the literature* for all researchers. To these factors we now add several more dynamics that will make the current system of publishing in the neurosciences even more untenable in the future.

First, neuroscience, as a discipline, has several characteristics that make the current system of publishing particularly problematic. The brain is a hugely complicated system, and its components cannot be easily studied in isolation, or strong conclusions drawn about the function of isolated components in the complete system. Progress depends on the development of large-scale theoretical frameworks and the building of consensus around the critical data that support, refine, or repudiate them. The intuitions and theories conveyed by a paper and the relationship between those theories and the literature are often as important as the data itself. The current system encourages novel seeming, isolated research, which is often directly contrary to establishing theories and interpretations in relation to the literature. Research designed to refine or address existing theories is relegated to specialist journals. This dynamic would be acceptable if this type of research was widespread, but there are few incentives to actually perform it. The lag and uncertainty in publication time and the relative uselessness of low-profile publications in promotion and tenure decisions rule out junior faculty or post-docs and these two groups perform most of the research in the labs of tenured faculty.

Second, the field of neuroscience, in both papers and researchers, is growing quickly. This year over 700 neuroscience doctorates will likely be awarded, compared with only 276 in 1993 (NSF Survey of Graduate Students). This increase in the number of researchers translates into an increase in the number of

submissions to existing journals (e.g., average annual increase from 2006 to 2009: nature 4.8%; Journal of Neuroscience 2.6%). The concomitant increase in the number of rejections and the ease of opening an online publication has also led to the creation of new journals. From 2000 to 2006 the number of neuroscience journals was essentially steady at around 200. From 2006 to 2009 the number of journals increased to 231, an annual increase of approximately 5% (derived from the Web of Science). As the field and the associated literature grow, the inefficiencies of the current system will become increasingly problematic. The amount of time it takes to publish a paper, the number of reviews written, and the difficulty in organizing and comprehending the literature will increase and eventually become a limiting factor on progress in the field, if it is not already.

Hopefully, this paper will help begin a conversation about the problems and inefficiencies inherent in the current system of publishing. The system proposed in this paper is not meant as a final proposal, but as a reasonable starting point that addresses many of the current flaws in the system and could reasonably be implemented. We hope that it will engender debate, which is at the heart of scientific progress, but too little emphasized in the current system of publishing.

This paper is also not meant to be an indictment of the existing journals; they are businesses whose purpose is to provide a service at a reasonable price. By and large they accomplish this purpose and are staffed by dedicated professionals wrestling with a difficult job. This paper is an indictment of the service that we, as a field, ask them to provide. We are paying, in both time and money, for a system constrained by the physical distribution of papers, when we no longer read physical copies of journals. What we should be paying for, and where private companies can be innovative, is in the coordination of the review process, the publicizing of results, and methods for searching and organizing the literature. Providing this last service can be quite profitable, Google has a profit margin of 21%. A better post-publication system will also improve the quality and frequency of scientific discussion between labs, which is now largely limited to conferences and published papers. In a time with increasingly constrained budgets and funding sources needing to see progress to justify taxpayer outlays, reforming the system of publishing might not only decrease our costs but increase our productivity as well.

## ACKNOWLEDGMENTS

Please note that the opinions expressed in this article are solely the opinion of the authors and do not necessarily reflect the opinion of the NIH. Thanks to Niko Kreigeskorte and Punitha Manavalan for many helpful discussions. Thanks to Sandra Truong for her extensive comments. Thanks to all those researchers who provided data for our analysis of the current system of peer review.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at [http://www.frontiersin.org/Computational\\_Neuroscience/10.3389/fncom.2011.00055/abstract](http://www.frontiersin.org/Computational_Neuroscience/10.3389/fncom.2011.00055/abstract)



## REFERENCES

- Fabiato, A. (1994). Anonymity of reviewers. *Cardiovasc. Res.* 28, 11–34.
- Lipworth, W. L., Kerridge, I. H., Carter, S. M., and Little, M. (2011). Journal peer review in context: a qualitative study of the social and subjective dimensions of manuscript review in biomedical publishing. *Soc. Sci. Med.* 72, 1056–1063.
- Peters, D. P., and Ceci, S. J. (1982). Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behav. Brain Sci.* 5, 187–195.
- Peters, D. P., and Ceci, S. J. (1984). How blind is blind review? *Am. Psychol.* 39, 1491–1494.
- Suls, J., and Martin, R. (2009). The air we breathe: a critical look at practices and alternatives in the peer-review process. *Perspect. Psychol. Sci.* 4, 40–50.
- UK House of Commons. (2004). *The Origin of the Scientific Journal and the Process of Peer Review*. Available at: <http://eprints.ecs.soton.ac.uk/13105/2/399we23.htm>
- Van Rooyen, S., Godlee, F., Evans, S., Black, N., and Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomized trial. *Br. Med. J.* 318, 23–27.
- Wright, G. (1994). Anonymity of reviewers. *Cardiovasc. Res.* 28, 1144–1144.
- Young, N. S., Ioannidis, J. P. A., and Al-Ubadli, O. (2008). Why current publication practices may distort science. *PLoS Med.* 5, e201. doi: 10.371/journal.pmed.0050201
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 June 2011; accepted: 11 November 2011; published online: 05 December 2011.

Citation: Kravitz DJ and Baker CI (2011) Toward a new model of scientific publishing: discussion and a proposal. *Front. Comput. Neurosci.* 5:55. doi: 10.3389/fncom.2011.00055

Copyright © 2011 Kravitz and Baker. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Alternatives to peer review: novel approaches for research evaluation

**Aliaksandr Birukou<sup>1,2\*</sup>, Joseph Rushton Wakeling<sup>2\*</sup>, Claudio Bartolini<sup>3</sup>, Fabio Casati<sup>1</sup>, Maurizio Marchese<sup>1</sup>, Katsiaryna Mirylenka<sup>1</sup>, Nardine Osman<sup>4</sup>, Azzurra Ragone<sup>1,5</sup>, Carles Sierra<sup>4</sup> and Aalam Wassef<sup>6</sup>**

<sup>1</sup> Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

<sup>2</sup> European Alliance for Innovation, Gent, Belgium

<sup>3</sup> Service Automation and Integration Lab, HP Labs, Palo Alto, CA, USA

<sup>4</sup> Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Catalonia, Spain

<sup>5</sup> Exprivia SpA, Molfetta, Italy

<sup>6</sup> Peerevaluation.org, Paris, France

## Edited by:

Diana Deca, University of Amsterdam, Netherlands

## Reviewed by:

Jelte M. Wicherts, University of Amsterdam, Netherlands

H. Steven Scholte, University of Amsterdam, Netherlands

Diana Deca, University of Amsterdam, Netherlands

Dietrich Samuel Schwarzkopf, Wellcome Trust Centre for Neuroimaging at UCL, UK

## \*Correspondence:

Aliaksandr Birukou and Joseph Rushton Wakeling, European Alliance for Innovation, Gent, Belgium, via alla Cascata 56/D, 38123 Trento, TN, Italy.  
e-mail: birukou@gmail.com;  
joseph.wakeling@gmail.com

In this paper we review several novel approaches for research evaluation. We start with a brief overview of the peer review, its controversies, and metrics for assessing efficiency and overall quality of the peer review. We then discuss five approaches, including reputation-based ones, that come out of the research carried out by the LiquidPub project and research groups collaborated with LiquidPub. Those approaches are alternative or complementary to traditional peer review. We discuss pros and cons of the proposed approaches and conclude with a vision for the future of the research evaluation, arguing that no single system can suit all stakeholders in various communities.

**Keywords:** research evaluation, peer review, metrics, bidding, opinions, LiquidPub, UCount

## 1. INTRODUCTION

Formal peer review of one kind or another has been part of the scientific publishing process since at least the eighteenth century (Kronick, 1990). While the precise norms and practices of review have varied extensively by historical period and by discipline (Burnham, 1990; Spier, 2002), key themes have remained consistent: a concern for ensuring the correctness of work and not allowing demonstrably false claims to distort the literature; the need for authors to have their work certified as valid; the reputation of the society, publisher, or editorial board responsible for the work; and at the same time, concern to not inhibit the introduction of valuable new ideas. Particularly with the increasing volume of publication through the twentieth century, the process has become an almost unavoidable necessity in determining what out of a huge range of submissions should be selected to appear in the limited (and costly) number of pages of the most prominent journals (Ingelfinger, 1974). One consequence of this competition for reader attention has been that reviewers are increasingly being asked to assess not just the technical correctness of work but also to make essentially editorial assessments such as the topical suitability and potential impact or importance of a piece of work (Lawrence, 2003).

Different practices for the evaluation of knowledge have been proposed and applied by the scientific community, including but not limited to *single-blind review* (where reviewers remain

anonymous, but author identity is known to the reviewer); *double-blind review* (where the identities of both authors and reviewers are hidden); and *open peer review* where both authors and reviewers are aware of each other's identity. Journal editors also have an important role, both in the initial assessment of whether to send a manuscript for review and in terms of management and final decision-making on the basis of reviewer recommendations; the precise degree of editor- versus reviewer-based selection can vary greatly between different publications (McCook, 2006). Yet despite its modern ubiquity, and a broad consensus among scientists upon its essential contribution to the research process (Ware and Monkman, 2008; Sense About Science, 2009), there are also widespread concerns about the known or perceived shortcomings of the review process: bias and inconsistency, ineffective filtering of error or fraud, and the suppression of innovation.

In this paper we discuss various models that offer complementary or replacement evaluation mechanisms to the traditional peer review process. The next section provides a brief overview of the conventional peer review process and its controversies, including a review of studies and analyses of peer review and reviewer behavior across a range of disciplines and review practices. This is followed by a review of a number of quantitative metrics to assess the overall quality and efficiency of peer review processes, to check the robustness of the process, the degree of agreement among and

bias of the reviewers, and to check the ability of reviewers to predict the impact of papers in subsequent years.

We then proceed to introduce a number of different experiments in peer review, including comparisons between quick ranking of papers, bidding to review papers, and reviewing them in the traditional manner. We also discuss two approaches to research evaluation that are based on leveraging on the explicit or implicit feedback of the scientific community: OpinioNet and UCount. We conclude the paper with a discussion of the pros and cons of the presented approaches and our vision for the future of the research evaluation.

## 2. PEER REVIEW HISTORY AND CONTROVERSIES

Review processes of one kind or another have been part of scientific publication since the first scientific journals – notably the *Philosophical Transactions* of the Royal Society – with the first formally defined peer review process being that of the journal *Medical Essays and Observations*, published in 1731 by the Royal Society of Edinburgh (Kronick, 1990). While historical practice varied greatly (Burnham, 1990), the growth of the scientific literature in the twentieth century has seen peer review become almost universal, being widely seen as the key evaluation mechanism of scholarly work (Ingelfinger, 1974; Ware and Monkman, 2008; Sense About Science, 2009).

Despite this ubiquity of the practice (or perhaps more properly, of a great diversity of practices coming under the same name), peer review has been little studied by scientists until the last decades. The results of these studies are perhaps surprising, being as they are often very equivocal about whether peer review really fulfills its supposed role as a gatekeeper for error correction and selection of quality work (Jefferson et al., 2007). A significant number of papers report that peer review is a process whose effectiveness “is a matter of faith rather than evidence” (Smith, 2006), that is “untested” and “uncertain” (Jefferson et al., 2002b), and that we know very little about its real effects because scientists are rarely given access to the relevant data.

For example, Lock (1994) claims that peer review can at most help detect major errors and that the real criterion for judging a paper is to look at how often its content is used and referred to several years after publication. Other experimental studies cast doubt on the ability of peer review to spot important errors in a paper (Godlee et al., 1998). At the same time, peer review is still considered a process to which no reasonable alternatives have been found (Kassirer and Campion, 1994; Smith, 2006).

Part of the problem is that the practice and goals of peer review can vary greatly by discipline and journal. Studies on peer review differ in the kind and amount of data available and use different metrics to analyze its effectiveness. Indeed, having precise objectives for the analysis is one of the key and hardest challenges as it is often unclear and debatable to define what it means for peer review to be effective (Jefferson et al., 2002a). Nevertheless, in general we can divide the metrics used into two groups: those aiming to determine the effectiveness or validity of peer review (discussed below), and those aiming at measuring what authors consider to be “good” *properties* of peer review (discussed in Section 3).

The first category of studies can itself broadly be divided into two categories: those testing the ability of peer review to detect errors, and those measuring reviewers’ ability to anticipate

the future impact of work, usually measured using citation count.

Where error detection is concerned, a study was conducted by Goodman et al. (1994) who studied 111 manuscripts submitted to the *Annals of Internal Medicine* between March 1992 and March 1993. They studied the papers before and after the peer review process in order to find out whether peer review was able to detect errors. They did not find any substantial difference in the manuscripts before and after publication. Indeed, they state that peer review was able to detect only small flaws in the papers, such as figures, statistics, and description of the results. An interesting study was carried out by Godlee et al. (1998), who introduced deliberate errors in a paper already accepted by the British Medical Journal (BMJ)<sup>1</sup> and asked 420 reviewers divided in 5 different groups to review the paper. Groups 1 and 2 did not know the identity of the authors, while 3 and 4 knew it. Groups 1 and 3 were asked to sign their reports, while 2 and 4 were asked to return their reports unsigned. The only difference between groups 4 and 5 was that reviewers from group 5 were aware that they were taking part in a study. Godlee et al. (1998) report that the mean number of major errors detected was 2 out of a total of 8, while there were 16% of reviewers that did not find any mistake, and 33% of reviewers went for acceptance despite the introduced mistakes. Unfortunately, the study does not report on whether the reviewers collectively identified all the errors (which might lend support to some of the community review processes discussed later in this article) or whether certain errors were noticed more often than others.

Citation count has been used extensively as a metric in studies by Bornmann and Daniel. The first of these reports on whether peer review committees are effective in selecting *people* that have higher citation statistics, and finds that there is indeed such a correlation (Bornmann and Daniel, 2005b). A later paper examines the initial assessments by staff editors of manuscripts submitted to a major chemistry journal, compared to the later assessments by external reviewers (Bornmann and Daniel, 2010a): where editors make an actual assessment this is indeed correlated with final citation count, but in 2/3 of cases they were unable or unwilling to venture an opinion. Final assessments after peer review were much more strongly correlated with final citation count, implying a positive effect whether or not editors were able to reach an initial decision. These results can be compared to those of Opthof et al. (2002) on submissions to a medical journal, where editors’ initial ratings were uncorrelated with later citation count, while external reviewers’ ratings were correlated, more strongly so where more reviewers were employed. The best predictive value, however, was a combination of reviewers’ and editors’ ratings, suggesting that differences in prediction ability are down to editors and reviewers picking up on different aspects of article quality.

## 3. QUANTITATIVE ANALYSES OF PEER REVIEW

In this section we review research approaches dealing with quantitative analysis of peer review. Effectiveness or validity of peer review can be measured taking into account different metrics,

<sup>1</sup>“With the authors’ consent, the paper already peer reviewed and accepted for publication by BMJ was altered to introduce 8 weaknesses in design, analysis, or interpretation” (Godlee et al., 1998).

included but not limited to: ability to predict the future position of the paper in the citation ranking, the disagreement between reviewers, the bias of a reviewer.

An obvious quantitative analysis is to measure the correlation between reviewers' assessments of manuscripts and their later impact, most readily measured by citation. As discussed in the previous section, results may be highly dependent on the particular context. For example, Bornmann and Daniel (2010b), studying a dataset of 1899 submissions to the *Angewandte Chemie International Edition*, found a positive correlation between reviewers' recommendations and the later citation impact – with, interestingly, a stronger correlation where *fewer* reviewers were used<sup>2</sup>. On the other hand, Ragone et al. (2011), studying a large dataset of 9000 reviews covering circa 3000 submissions to 10 computer science conferences, observed few statistically significant correlations when the ranking of papers according to reviewer ratings was compared to the ranking according to citation<sup>3</sup>.

Another important metric for the peer review process is the inter-reviewer agreement (Casati et al., 2010), which measures how much the marks given by reviewers to a contribution differ. The rationale behind this metric is that while reviewers' perspectives may differ according to background, areas of expertise and so on, we may expect there to be some degree of consensus among them on the core virtues (or lack thereof) of an article. If on the other hand the marks given by reviewers are comparable to marks given at random, then the results of the review process are also effectively random, which defeats its purpose. There are several reasons for having several reviewers per contribution: to evaluate based on consensus or majority opinion and to provide multiple expertise (e.g., having a more methodological reviewer and two more content reviewers).

Indeed, having a high disagreement value means, in some way, that the judgment of the involved peers is not sufficient to state the value of the contribution itself. This metric could be useful to improve the quality of the review process as could help to decide, based on the disagreement value, if three reviewers are enough to judge a contribution or if more reviewers are needed in order to ensure the quality of the process.

A significant portion of the research on peer review focuses on identifying reviewer *biases* and understanding their impact in the review process. Indeed, reviewers' objectivity is often considered a fundamental quality of a review process: “the ideal reviewer,” notes Ingelfinger (1974), “should be totally objective, in other words, supernatural.” Approaches for analyzing bias in peer reviews identified several kinds of bias: *affiliation* bias, meaning that researchers from prominent institutions are favored in peer review (Ceci and Peters, 1982); bias in favor of US-based researchers (Link, 1998),

*gender* bias against female researchers (Wenneras and Wold, 1997; Bornmann, 2007; Marsh et al., 2009; Ceci and Williams, 2011) and *order bias* (Bornmann and Daniel, 2005a), meaning that reviewing applications for doctoral and post-doctoral research scholarship in alphabetic order may favor those applicants having names at the beginning of the alphabet. Although it is not always easy to decouple these apparent biases from other factors such as quality differentials, at least some biases, such as those based on nationality of reviewers and authors, remain even when quality is taken into account (Lee et al., 2006; Lynch et al., 2007). Others, such as bias in favor of statistically significant results (Olson et al., 2002; Lee et al., 2006) or gender biases (Marsh et al., 2009; Ceci and Williams, 2011), appear to be down primarily to other factors than the review process itself. In addition, it is possible to compute the *rating bias*, i.e., reviewers consistently giving higher or lower marks, independently from the quality of the specific contribution they have to assess, which is a kind of bias that appears rather often, is easy to detect, and that can be corrected with rather simple procedures to improve the fairness of the review process (Ragone et al., 2011).

One of the ways to identify bias is to compare single- and double-blind review. Single-blind review provides anonymity to the reviewers and is used to protect the reviewers from author reprisals. In many research fields, single-blind review is the normative practice. However, in others, such as information systems, or at Association for Computing Machinery Special Interest Group on Management of Data (ACM SIGMOD) conferences, double-blind review, where identities of both authors and reviewers are hidden, is the norm. The purpose of the double-blind review is to help the reviewers to assess only scientific achievements of the paper, not taking into consideration other factors and therefore to be unbiased.

Analyses of the merit of the double-blind review process are somewhat equivocal. Early studies by McNutt et al. (1990) and Fisher et al. (1994) on double-blind review of journal submissions reported a positive effect on review quality as rated by editors, although the latter study may have been influenced by the fact that blinded reviewers knew they were taking part in a study while non-blinded reviewers did not. A later and much larger study by Justice et al. (1998), where all reviewers knew they were taking part in a study, revealed no statistically significant difference, while another by van Rooyen et al. (1999) including both informed and uninformed reviewers suggested no difference due to either the review style (single- or double-blind) or reviewer knowledge of whether they were partaking in a study. On the other hand an extensive study of abstract submissions to medical conferences by Ross et al. (2006) suggested that double-blind review was successful in eliminating a host of biases related to gender, nationality, prestige, and other factors.

One major factor that may explain these contradictory results is the question of whether the masking of author identity is actually successful, as authors frequently include identifying elements in their papers such as citations to their previous work (Cho et al., 1998; Katz et al., 2002). The likelihood of such accidental unblinding may be larger for extensive works like journal submissions, making it more difficult for double-blind review to succeed compared to shorter works such as abstracts. Unblinding rates vary widely between journals, and it may be that volume of submissions

<sup>2</sup>This marks an odd contradiction to the results of Opthof et al. (2002), where more reviewers made for better prediction. One explanation might be that in medical research there could be a greater number of different factors that must be considered when assessing an article, hence several reviewers with different expertise might produce a better review.

<sup>3</sup>Correlation between reviewer- and citation-based rankings was measured using Kendall's  $\tau$  for 5 different conferences, of which 2 had weak but statistically significant correlations ( $\tau = 0.392$ ,  $p = 0.0001$  and  $\tau = 0.310$ ,  $p = 0.005$ ; the two conferences had respectively 150 and 100 submissions). The other, larger conferences had no statistically significant correlation (Mirylenska et al., unpublished).

and the size of the contributing community also affect how easy it is to identify authors (Ross et al., 2006). It may also be possible for authors to identify reviewers from their comments. Potential positive effects of successfully blinded review may therefore be difficult to secure in practice.

Research on open peer review (where the reviewer's name is known to the authors) is at present very limited. Initial studies showed that open reviews were of higher quality, were more courteous and reviewers spent typically more time to complete them (Walsh et al., 2000). An example of the open peer review, adopted mainly by \*PloP<sup>4</sup> conferences, is *shepherding*, where a shepherd (reviewer) works together with the sheep (authors) on improving the paper. The major problem of open peer review is combating the unwillingness of some potential reviewers to agree to their identity being revealed (Ware and Monkman, 2008), although journals that have implemented open review have reported good experiences in practice (Godlee, 2002).

Research shows that to improve the peer review process, sometimes paying attention to details is enough. For instance, the mark scale can influence reviewers and lead them to use only specific marks, instead of the whole scale (Casati et al., 2010; Medo and Wakeling, 2010). It has been shown that in the scale from 1 to 5 with half-marks, reviewers tend to not use half-marks, while in the same scale without half-marks (1 to 10) reviewers use the entire scale to rate (Casati et al., 2010). In a scale from 1 to 7, reviewers' marks tend to concentrate in the middle (Casati et al., 2010).

One of the main issues in peer review analysis is to have access to the data. Usually, works on peer review are restricted to analyzing only 1–2 conferences, grant applications processes or fellowships. Just to name a few: Reinhart (2009) analyzed 496 applications for project-base funding; Bornmann and Daniel (2005a) studied the selection process of 1,954 doctoral and 743 post-doctoral applications for fellowships; Bornmann et al. (2008) analyzed 668 applications for funding; Godlee et al. (1998) involved in their experiments 420 reviewers from the journal's database; Goodman et al. (1994) analyzed 111 manuscripts accepted for publication. As already mentioned above, one of the largest datasets has been used in the work by Ragone et al. (2011) where they collected data from 10 conferences, for a total of 9032 reviews, 2797 submitted contributions and 2295 reviewers.

#### 4. EXPERIMENTS IN PEER REVIEW

Nowadays, scientists and editors are exploring alternative approaches to tackle some of the pervasive problems with traditional peer review (Akst, 2010). Alternatives include enabling authors to carry reviews from one journal to another (Akst, 2010), posting reviewer comments alongside the published paper<sup>5</sup>, or running the traditional peer review process simultaneously with a public review (Akst, 2010). The ACM SIGMOD conference has also experimented with variations of the classical peer review model where papers are evaluated in two phases, where

the first phase filters out papers that are unlikely to be accepted allowing to focus the reviewers' effort on a more limited set of papers. In Casati et al. (2010) authors provide a model for multi-phase review that can improve the peer review process in the sense of reducing the review effort required to reach a decision on a set of submitted papers while keeping the same quality of results.

In the following we focus on three experimental approaches for peer review: asking reviewers to rank papers instead of reviewing them, bidding for reviewing a paper, and open evaluation of research works.

##### 4.1. EXPERIMENT ON RANKING PAPERS vs REVIEWING

For the Institute of Electrical and Electronics Engineers Business-Driven IT Management Workshop (IEEE BDIM) in 2010, the Technical Program Committee (TPC) chairs experimented with a "wisdom of the crowd" approach to selecting papers. The aim of the experiment was to assess the viability of an alternate selection mechanism where (some of the) reviewers can rank papers based on a quick read rather than providing an in-depth review with quality scores.

This is the process they followed:

- The TPC members were asked to split into two roughly equal-size groups: (a) "*wisdom of the crowd*" and (b) "*traditional*," TPC chairs completed the split for those TPC member who did not reply or were indifferent<sup>6</sup>.
- TPC members obviously knew which group they were in, but had no direct knowledge of other members' placement.
- Group (b) carried out the usual 3–4 traditional reviews.
- At the end of the review phase, reviews from group (b) were averaged as usual, resulting in a total order of all papers submitted.
- Group (a) got assigned a PDF containing all submissions (excluding conflicts of interest) *with no author information*, thus we followed double-blind review process.
- Group (a) was required to provide a total order of all (or most) of the papers submitted, spending no more than 3–5 min on each paper.
- They TPC chairs merged the lists giving equal weight to each, and the top papers were divided into tiers (extended presentation, regular presentation, short presentation, posters, rejected) according to the harmonized ordered list. TCP chairs performed tie-break where necessary.
- Authors received feedback containing
  - Acceptance/rejection;
  - Tier of acceptance if applicable (extended, regular, short, poster);
  - Full explanation of the review process;
  - at least 3 reviews for their submission;
  - their paper's rank in the traditional review process, and its rank in the "wisdom of the crowd" process.

<sup>4</sup>PloP stands for Pattern Languages of Programs and \*PloP family of conferences includes: EuroPloP, PloP, VikingPloP, etc. See <http://www.hillside.net/europlop/europlop2011/links.html> for a complete list.

<sup>5</sup><http://interdisciplines.org/>, a website for interdisciplinary conferences run as conversations.

<sup>6</sup>Note that technically this experiment is closer to a quasi-experiment because the reviewers were allowed to choose the type of review process. If any of the groups was superior in terms of reviewing quality, this may have affected the results.

Interesting findings were:

- (1) reviewers split evenly between the two groups, with exactly half of the TPC choosing the “wisdom of the crowd” approach, and half choosing the traditional
- (2) for selection, there were three traditional reviewers for each paper, so the TPC chairs counted the score from the wisdom of the crowd ranking with a weight equal to three reviewers. They transformed the ranking into a score by averaging ranks over all the reviewers, and normalizing linearly the average rank onto the range of scores of the traditional reviews
- (3) results were such that the top three papers and the bottom four papers were identical for both the traditional and the fast ranking review. However, for the selection of the papers in the middle, the TPC chairs had to take into account not only review scores, but also the review content, and give more weight to more experienced reviewers. For the submissions falling in the in-between category, the wisdom of the crowd did not help, and it was mostly off what the end selection wound up being.

In conclusion, the experiment showed that fast ranking in the wisdom of the crowd approach could be applied to get to a fast selection of the top and bottom submissions. However, that does not help in selecting the papers that fall in-between these categories.

## 4.2. e-SCRIPTS: BIDDING FOR REVIEWING

Most researchers maintain a strong preference for peer review as the key mechanism of research evaluation (Ware and Monkman, 2008; Sense About Science, 2009). A major motivating factor here is the ability of peer review not just to assess or filter work but to help *improve* it prior to publication (Goodman et al., 1994; Purcell et al., 1998; Sense About Science, 2009), and many researchers consider this opportunity to help their fellow scientists to be one of the key pleasures of contributing reviews (Sense About Science, 2009).

By contrast, some of the major frustrations of authors (and editors) with the review process relate to those occasions when the reviewer is unmotivated or unfamiliar with the subject matter. At conferences (e.g., at EuroPLOP), this factor is often dealt with by allowing members of the technical program committee to *bid* to review submissions on the basis of titles and abstracts. In this way, every program committee member can hope to have a paper to review which meets their interests and areas of expertise. The role of the program chair is also made easier, with less work to do in assigning referees to articles.

The *e-Scripts* submissions management system<sup>7</sup>, developed by the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (ICST) and the European Alliance for Innovation (EAI), attempts to bring the same principles and benefits to the peer review process of research journals. Titles and abstracts of submitted articles are posted publicly online after submission, and for a period of about 2 weeks thereafter interested readers can bid to review those which catch their fancy. At the

end of the public bidding period, the editor approves an ordered list of candidate reviewers based on a mix of bidders, author- and editor-nominated candidates, and reviewer invitations are sent out automatically starting from the top of the list.

The aim here is principally to engage with the enthusiasm and willingness to help that motivate good reviewers, while not relying on it: as opposed to some unsuccessful attempts at community review (Greaves et al., 2006), the Editors still have a responsibility to nominate and secure reviewers, with bidding acting as a supplemental rather than replacement selection process. In addition the system maintains a level of confidentiality for unpublished work, with the journal Editor still controlling access.

Beyond improving the quality of individual reviews, this approach has the capacity to generate additional data to support editorial decision-making. First, just as early download statistics offer a reliable precursor of later citation impact (Brody et al., 2006), so we can anticipate bidding intensity to reflect the potential importance of a submitted article. Second, correlations in user bidding can be used to build a profile of reviewer interests that can help automate the process of reviewer nomination. This, together with other means of assessing and ranking potential reviewers, is the subject of EAI's *UCount* project, which is discussed in Section 5.2.

## 4.3. PEEREVALUATION.ORG: SCIENTIFIC TRUST IN THE SOCIAL WEB

For the Millennial generation, sharing, reviewing, disseminating, and receiving immediate feedback have become not only natural practices but also strong expectations. For almost a billion Facebook users, both practices and expectations are fully embedded in the daily flows of consumption, communication, entertainment, information, work, and access to knowledge.

### 4.3.1. The advent of social reputation

On the Social Web, all are empowered to become, all at once, producers, reviewers, disseminators, and consumers. With such empowerment and shuffling of roles, it is only logical that alternative mechanisms of *reputation building* would also emerge.

### 4.3.2. The story of John

John composed a song, uploaded it on YouTube and sent it to his friends. The song became a hit and triggered exponential viral dissemination. John has now a reputation as a composer and has built a network of 500,000 thousand listeners, fans, and reviewers. In John's story, music publishers, distributors, and journalists had no implications in the realization of his endeavors. John relied on social dissemination, reviewing, and *social reputation building*. He was then offered a contract by a music label, which he chose to accept, for greater dissemination and recognition.

### 4.3.3. The story of Sophie

John's younger sister, Sophie, is a neurobiologist who just defended her Ph.D. Sophie is as Web savvy as John and expects her career to be just as fluid. Sophie knows that her future as a researcher will depend on her capacity to contribute to neurobiology with original and valid methods and results, and sufficient funding. To convince research funding agencies, all that Sophie needs is a method to certify that her research projects are valuable to neurobiology and that her methods and results are valid. Sophie is

<sup>7</sup><http://escripts.icst.org/>



of course aware that she could publish articles in peer reviewed journals to give tokens of trust to such agencies but, having knowledge of John's experience, she is disappointed by the slowness of the peer reviewing process, publishing costs and the complex and opaque mechanisms of scientific reputation and impact measures. Indeed, like John, Sophie values empowerment, immediacy, transparency, and qualitative appreciation of her work, as opposed to automated and quantitative measures of her impact.

#### 4.3.4. Sophie's world

Sophie does not need 500,000 viewers or reviewers. In her smart-phone, she has the email addresses of 20 peers around the World specializing in her field, 20 neurobiologists who could review her work. All she needs is a place where she can demonstrate that she has respected the rules of *scientific trust* and that her methods and results have indeed been reviewed by qualified and objective peers. This place should also be *social dissemination friendly* so that her work may be shared, discussed and recommended by an exclusive community of specialized peers.

Finally, because research funding agencies are usually overwhelmed by the number of proposals, Sophie will have to provide them with a summarized and comprehensive digest representing to what extent her research is indeed valid, original and endorsed by peers who believe it is useful to science, and to human development at large.

These are the issues [peerevaluation.org](http://peerevaluation.org) is tackling all at once, aware that a platform supporting Open Science, collaborative peer reviewing and dissemination cannot succeed without powerful incentives, innovative intellectual property rights management and, finally, reliable representations of scientific trust that meet the expectations of policy makers and funding bodies.

Peerevaluation.org aims at becoming a place where scholars come to make sure that they are getting the best of online sharing: increased dissemination, visibility, accessibility, commentary, and discussion, fruitful collaborations and, finally, evidence of impact, influence and re-use.

The basic peerevaluation.org scenario – focusing on the dissemination and remote pre- or post- publication peer review and commentary – unfolds as follows: (a) you upload a PDF of your recent paper; (b) you export the PDF's abstract and link to your blog, your Mendeley account and a repository like CiteSeerX. (c) simultaneously it gets indexed by Google Scholar and Microsoft Academic Search; (d) wherever your file is, people can comment it, discuss it, recommend it, share it, have access to your articles statistics, social impact measures; (e) all these remote social interactions are simultaneously aggregated and displayed in your [peerevaluation.org](http://peerevaluation.org) account, for you and others to consult.

## 5. APPROACHES FOR COMMUNITY-BASED EVALUATION

Existing problems in peer review and new tools brought by Web 2.0 triggered new directions in research evaluation, making trust and reputation an important topic for peer review (see, for instance, the [Peerevaluation.org](http://Peerevaluation.org) approach). Reputation reflects community opinion on the performance of an individual with respect to one or more criteria. In this section we review two approaches for research evaluation leveraging on the explicit or implicit feedback of the scientific community, namely: (1) OpinioNet computes the

reputation of researchers based on the opinions, such as review scores or citations; (2) UCount employs dedicated surveys to elicit community opinion on individual's performance either as a researcher, or as a reviewer.

### 5.1. OPINIONET: REPUTATION OF RESEARCH BASED ON OPINION PROPAGATION

OpinioNet is a tool that is based on the notion of the propagation of opinions in structural graphs. In OpinioNet, the reputation of a given research work is not only influenced by the opinions it receives, but also by its position in the publications' structural graph. For instance, a conference is reputable because it accepts high quality papers. Similarly, people usually assume that in the absence of any information about a given paper, the fact that the paper has been accepted by a highly reputable journal implies that the paper should be of good quality. Hence, there is a notion of propagation of opinions along the *part\_of* relation of structural graphs.

Figure 1 provides an example of a common structural graph of research work. In this figure, there is a conference series CS that has a set of conference proceedings,  $\{CP_1, \dots, CP_n\}$ , and each conference proceeding is composed of a set of papers. Similarly, there is a journal J that has a set of volumes,  $\{V_1, \dots, V_n\}$ , each composed of a set of papers. We note that if papers were split into sections,  $\{S_1, \dots, S_m\}$ , then it is possible for different papers to share some sections, such as the "Background" section.

Current reputation measures in the publications field have mainly focused on citation-based metrics, like the *h*-index. Explicit reviews (or opinions) have been neglected outside the review process due to the fact that this information is very scarce in the publications field, unlike e-commerce scenarios such as Amazon or eBay. OpinioNet addresses this problem by providing means that help a single researcher infer their opinion about some research work (or other researcher) based on their own opinions of bits and pieces of the global publications structural graph. Accordingly, the reputation (or group opinion) is calculated by aggregating individual researchers' opinions.

Furthermore, OpinioNet may also be used with indirect opinions. When computing the reputation of researchers and their research work, we say there is a lot of information out there that may be interpreted as opinions about the given researcher or research work. For instance, the current publication system provides us with direct (explicit) opinions: the review scores.

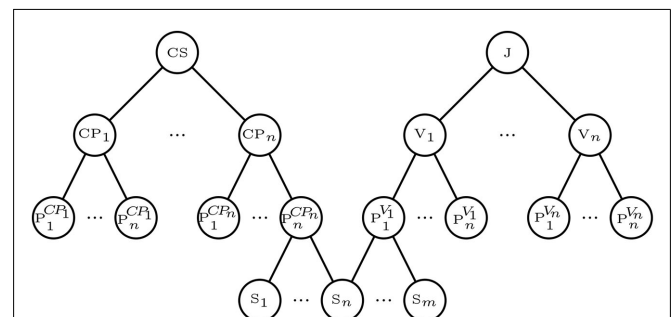


FIGURE 1 | A sample structural graph in the publications field.

Additionally, direct (or implicit) opinions may also be considered. For example, citations may also be viewed as an indication of how good a given research work is, i.e., a positive opinion of the citing authors about the cited research work. Subscription to journals may be viewed as an indication of how good a journal is viewed in its community, i.e., a positive opinion of the subscriber about the journal. Massive volumes of information exist that may be interpreted as opinions. The OpinioNet algorithm (Osman et al., 2010b) uses these opinions, whether they were direct or indirect, to infer the opinion of a researcher about some given research work<sup>8</sup>, and then infer the opinion of a research community accordingly. More importantly, OpinioNet may be used for any combination of information sources, although different fields of research may give more weight to one information source over the other.

As such, OpinioNet is easily customizable to suit the requirements of different communities or disciplines. For example, it is known that different disciplines have very different traditions and attitudes toward the way in which research is evaluated. With OpinioNet, one can select the source(s) of opinions to focus on, possibly giving more weight to different sources. For instance, one may easily make OpinioNet run on one's own personal opinions only, the direct opinions of the community, on citation-based opinions only, or on a combination of citation-based opinions and direct ones. OpinioNet may also give more weight to papers accepted by journals that conferences, or *vice versa*. And so on.

Furthermore, OpinioNet does not need an incentive to encourage people to change their current behavior. Of course, having an open system where people read and rate each others work would be hugely beneficial. But OpinioNet also works with the data which is available now. We argue that we already have massive numbers of opinions, both direct and indirect, such as reviews, citations, acceptance by journals/conferences, subscriptions to journals, references from untraditional sources (such as blogs), etc. What is needed is a system, such as OpinioNet that can access such data, interpret it, and deduce reputation of research work accordingly. At the time being, we believe that accessing and compiling this data is the main challenge.

As for potential bias, when considering an opinion, the reputation of the opinion source is used by OpinioNet to assess the reliability of the opinion. For example, we say a person that is considered very good in a certain field is usually considered to be very good as well in assessing how others are in that field. This is based on the *ex cathedra* argument. An example of a current practice following the application of this argument is the selection of members of committees, advisory boards, etc. Although, of course, instead of simply considering the expertise of the person in the field, complementary methods that may assess how good the person is in rating research work may be used to enrich OpinioNet against bias and attacks. For example, studying a person's past reviews could tell whether the person is usually biased for a specific gender, ethnicity, scientific technique, etc. Also, analyzing past reviews, one may also tell how close a person's past opinions were to the group's opinion. Past experiences may also be used to

assess potential attacks, such as collusion. All of this information is complementary to OpinioNet, and it may be used by OpinioNet to help determine the reliability of the opinion.

After introducing the basic concepts and goals of OpinioNet, we now provide a brief technical introduction to the algorithm. Of course, for further details, we refer the interested reader to Osman et al. (2010b). And for information about evaluating OpinioNet and its impact on research behavior via simulations, we refer interested readers to Osman et al. (2011).

### 5.1.1. Reputation of research work

The reputation of research work is based on the propagation and aggregation of opinions in a structural graph. OpinioNet's propagation algorithm is based on three main concepts:

- *Impact of a node.* Since researchers may write and split their research work into different 'child nodes' (e.g., a section of a paper, or papers in conference proceedings), it is impossible to know what is the exact weight to assign to each child node when assessing its impact on its parent nodes (and *vice versa*). In OpinioNet, the impact of a given node  $n$  at time  $t$  is based on the proportion of nodes that have received a direct opinion in the structural sub-tree of  $n$ . In other words, OpinioNet relies on the attention that a node receives (whether positive or negative) to assess its impact. For example, if one paper of a journal received a huge number of reviews (positive or negative) while another received no attention at all, then the one that received a huge number of reviews will have a stronger impact on the reputation of the journal than the latter.
- *Direction of propagation.* The direction of propagation in the structural graph is crucial. Each holds a different meaning. The "downward" propagation is viewed to provide the *default* opinion, such as a paper inheriting the reputation of the journal that accepted it. The default opinion is understood to present the opinion about the node that is inherited from the parents, and is usually used when there is a lack of information about the children nodes that help compose the node in question. The "upward" propagation provides the *developing* opinion, such as a conference aggregating the reputation of its papers. Then, each time a new opinion is added to a node in the graph, the default and developing opinions of its neighboring nodes are updated accordingly. Then, the update of one node's values triggers the update of its neighboring nodes, resulting in a propagation wave throughout the structural graph.
- *Decay of information value.* We say everything loses its value with time. Opinions are no exception, and an opinion about some node  $n$  made at time  $t$  loses its value (very) slowly by decaying toward the decay probability distribution (or the default opinion) following a *decay function* that makes the opinion converge to the default one with time.

We note that OpinioNet essentially propagates the opinions of *one* researcher on a given attribute (say quality of research) in a structural graph. However, opinions may be provided for several attributes, such as novelty, soundness of research work, etc. Opinions may also be provided by more than one researcher. In these cases, different aggregations may be used to obtain the final

<sup>8</sup>How indirect opinions may be defined is an issue that has been addressed by Osman et al. (2010a).



group opinion about a given piece of research work. Osman et al. (2010b) provides some examples on how to aggregate these opinions to obtain a final reputation measure. However, as discussed earlier, an important thing to note is that the reputation of each opinion holder is used to provide a measure on how reliable their opinions are. In other words, the reputations of opinion holders are used to provide the weights of the opinions being aggregated.

### 5.1.2. Reputation of researchers

Every node of a structural graph has its own author, or set of coauthors. The authors of different sections of a paper may be different, although there might be some overlap in the sets of authors. Similarly, the authors of different papers of a conference may be different. And so on. In OpinioNet, the reputation of an author at a given time is an aggregation of the reputation of its research work. However, the aggregation takes into consideration the number of coauthors that each paper has. The aggregation (see Osman et al., 2010a) essentially states that the more coauthors some research work has, the smaller the impact it leaves on each of its coauthors.

## 5.2. UCOUNT: A COMMUNITY-BASED APPROACH FOR RESEARCH EVALUATION

The UCount approach<sup>9</sup> (Parra et al., 2011) provides the means for community-based evaluation of overall scientific excellence of researchers and their performance as reviewers. The evaluation of overall scientific excellence of researchers is done via surveys<sup>10</sup> that aim at gathering community opinions on how valuable a given

researcher's contribution to science is. The results are aggregated to build rankings. In the current section we describe the use of UCount for assessing reviewers, since it better fits the scope of the special issue.

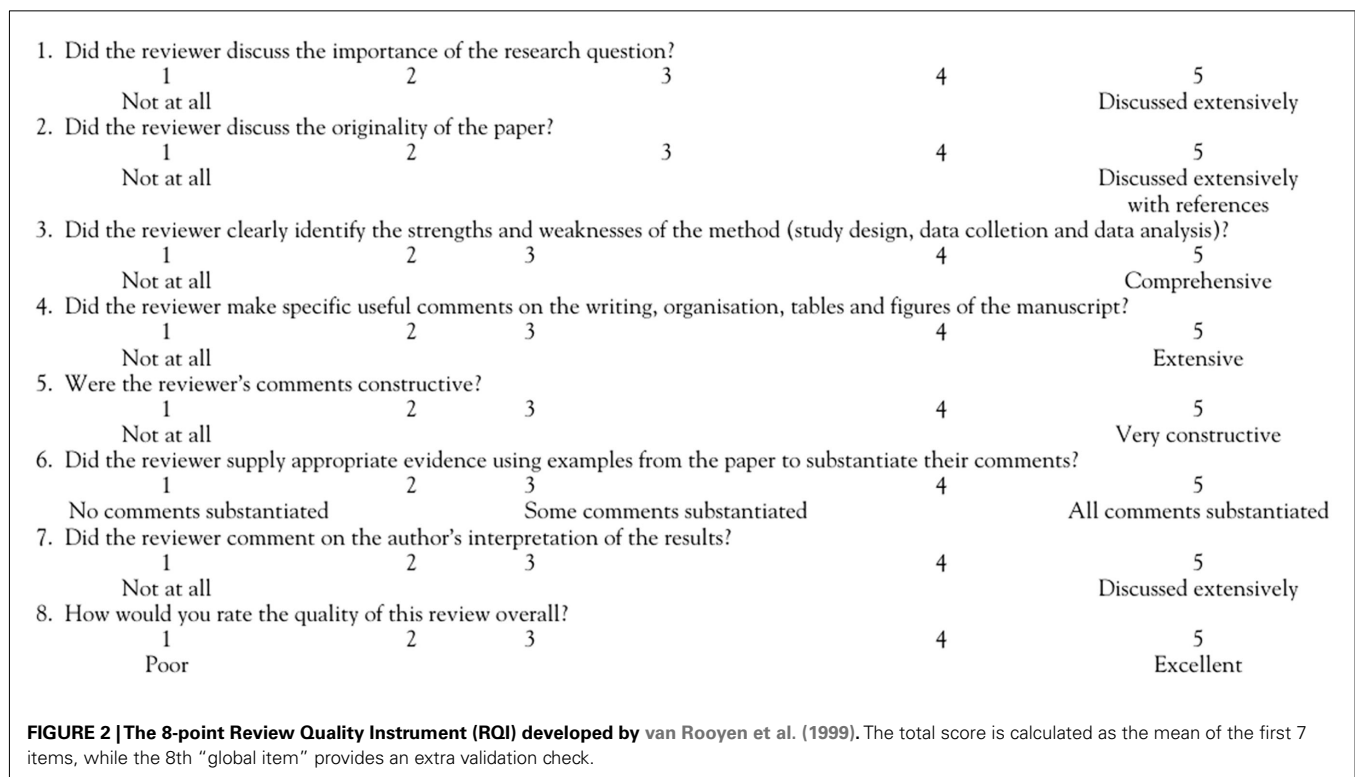
UCount for assessing reviewers is specifically designed to operate based on reviewer performance *as reviewers*, as opposed to other criteria such as bibliometric prominence: a high-profile researcher is not necessarily a good reviewer (Black et al., 1998). UCount is integrated in the above-mentioned e-Scripts, a review system used for the ICST Transactions. It enables authors and editors to provide feedback on the performance of reviewers using the Review Quality Instrument (RQI) developed by editors of the British Medical Journal (van Rooyen et al., 1999). This is a psychometrically validated instrument used in multiple studies of peer review (Jefferson et al., 2007).

The RQI consists of an 8-point scale (Figure 2), where each item is scored on a 5-point Likert scale (1 = poor, 5 = excellent). The first 7 points each enquire about a different aspect of the review, including the discussion of the importance and originality of the work, feedback on the strengths and weaknesses of the research method and the presentation of the results, the constructiveness of comments, and the substantiation of comments by reference to the paper. The 8th and final item is an overall assessment of the review quality, and can be compared to the total score calculated as the mean of the first 7 items.

On the basis of this feedback, every 3 months (linked with ICST Transactions issue schedule) public rankings of reviewers will be presented. Reviewers submitting at least three reviews will be ranked according to several criteria: overall best score, total number of reviews completed, and the usefulness, insight, and

<sup>9</sup><http://icst.org/ucount/>

<sup>10</sup>See examples of such surveys at <http://icst.org/UCount-Survey/>



constructiveness of feedback. Moreover, during the process of choosing the reviewers for a paper, the editor will be able to see the ranking of the reviewers based on their past performance. The ranking will be based on RQI feedback:

- First-placed are candidates with a mean RQI score higher than a given threshold (suggest the median 3), ranked according to their RQI score.
- Next come candidates with no RQI, including both new reviewers and those who have completed less than 3 reviews in the last 12 months. These candidates will be ranked in the traditional bidder-author-editor order.
- Last come candidates whose mean RQI score is *below* the acceptable threshold, ranked in descending order of score.

Where available, RQI for candidates will be displayed in order to clarify the ranking. Editors will still be able to re-order the candidate list. We believe that this will lead to the selection of better reviewers and also to their recognition in the community as opposed to the current situation in most journals, where only the members of the editorial board get credits, while the reviewers remain unknown.

UCount is now being implemented for publication activities of the European Alliance for Innovation (EAI) and the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (ICST).

## 6. CONCLUSION AND DISCUSSION

In this paper we have presented a range of possible extensions or alternatives to the conventional peer review process. The diversity of these approaches reflects the wide range of complementary factors that can be considered when determining the value of a scientific contribution. Indeed, definitions of quality are often highly context-dependent: for example, in some cases a technically unreliable but imaginative and inspirational paper may be of more value than a thorough and careful examination (Underwood, 2004), while in other cases, the opposite will be true. Such a diversity of needs requires a diversity of solutions.

The particular selection of the approaches for research evaluation reviewed in this paper is by no means complete, reflecting primarily the research carried out by the LiquidPub<sup>11</sup> project and its collaborators<sup>12</sup>. There exist many other approaches that we would see as complementary, for example expert expert post-publication review such as that carried out by the Faculty of 1000<sup>13</sup>, or personalized recommender systems (Adomavicius and Tuzhilin, 2005; Zhou et al., 2010).

In the following we discuss controversial aspects of the approaches reviewed in the paper.

### 6.1. BIDDING AS AN INDICATOR OF IMPORTANCE

Given the known results regarding article download statistics (Brody et al., 2006) and the findings from the experiment on bidding described in Section 1, we can expect that bid counts too will

serve as a reliable (though not infallible) indicator of the future impact of research work. A concern here is that – as with citation – people may bid not just on papers which interest them topically, but on papers which they wish to criticize and see rejected. Our inclination is that this is less of a risk than might be thought, for two main reasons. First, results from online rating systems such as the 5-star system used on YouTube show that there is a very strong bias toward positive ratings, suggesting that people treat items which they dislike with indifference rather than active criticism (Hu et al., 2009): we can expect that a similar principle may apply in bidding, that potential reviewers will ignore bad papers rather than waste valuable time volunteering to critique something they will likely expect to be rejected anyway. Second, leaving aside bad papers, we may anticipate bidders volunteering to review papers with which they have a strong disagreement. This may certainly create an issue for the journal Editor who must control for the potential conflicts of interest, but it does not reflect a conflict with the potential impact of the paper. Papers on hotly contested topics are likely to be more, not less, highly cited.

An additional risk is that since bidding is based on title and abstract, it may attract attention to “over-sold” papers whose claims are made to sound more important than they actually are. This is of course a universal problem of research, not limited to bidding: authors try and over-hype their work to attract editorial, reviewer, and reader attention (Lawrence, 2003). The major question, which will have to be addressed on the basis of future experience, is whether this will distort the bidding statistics any more than it already does the citation and download counts.

On a more positive note, bidding is in line with one of the key motivations for scientists to engage in peer review, namely that by doing so they can help to improve and contribute to their colleagues’ work (Goodman et al., 1994; Purcell et al., 1998; Sense About Science, 2009). This strong ethic of professional altruism is more than likely to help offset the risks described above, and provides another reason why bidding is likely to reflect importance and impact – it is more exciting to contribute to work which you believe will be of lasting importance.

### 6.2. PEEREVALUATION.ORG vs. UCOUNT

Peerevaluation.org and UCount both aim at more open and transparent peer review. However, while UCount aims at incremental change in the traditional journal review, by introducing feedback on the reviewers, Peerevaluation proposes a radical shift in the process, which in its case is no more managed by the editors. We believe that the two approaches can be combined in the future, for instance UCount findings can be used to suggest reviewers in Peerevaluation, while Peerevaluation past review history can be a valuable input to UCount.

### 6.3. USE OF COMMUNITY OPINIONS

OpinioNet and UCount approaches use community opinions to estimate the reputation of a researcher. To take into account that majority is not always right, OpinioNet weights opinions based on the credibility of the opinion source, e.g., the level of expertise of the person who provides the opinion. UCount, however, aims at catching the community opinion as it is, without any adjustments. Therefore, UCount does not aim at answering “is it true that person

<sup>11</sup><http://project.liquidpub.org/>

<sup>12</sup>A complete overview of the research carried out by the project on these topics is available at <http://project.liquidpub.org/research-areas/research-evaluation>

<sup>13</sup><http://f1000.com/>

A is the best reviewer (researcher)?”, but rather at stating “community X thinks that person A is the best reviewer (researcher).” Both approaches rely on getting data about community opinions: while OpinioNet aims at collecting the data already available via citation, co-authorship, and publication networks, UCount requires that authors fill in a questionnaire, and the results can be used as direct opinions in OpinioNet.

#### 6.4. INCENTIVES TO PARTICIPATE

Providing direct opinions on reviewers in UCount might be seen as yet another action required from the author. However, providing ratings is a minimal effort comparing to writing a paper or writing a review. Therefore, we believe that if really good journals will require feedback on reviewers (e.g., as proposed by UCount), then people will participate and then other journals will have to follow. Moreover, in both the UCount and Peerevaluation approaches reviewers have incentives to submit good reviews because they know they are being assessed, either directly (UCount) or indirectly (Peerevaluation, because reviews are public). Moreover, reviewers will get publicity for doing a good job. UCount also offers incentives for authors, who are encouraged to participate because in this way they help editors to select better reviewers, and therefore, get better reviews. If at some point in time it appears that there are not enough good reviewers, maybe the incentives should be reconsidered. Controversial but possible incentives include paying reviewers, making it possible to submit a paper only after first reviewing three other papers, or reducing registration fees for people who spend time reviewing papers for a conference.

#### 6.5. THE ROLE OF THE INTERNET

It has long been recognized that the advent of the Web offers many opportunities to change the landscape of research publication and evaluation (Harnad, 1990; Ginsparg, 1994; Swan, 2007). At the most basic level, electronic publication effectively reduces storage, distribution, and communication costs to near zero, as well as greatly facilitating the creation and sharing of documents (Odlyzko, 1995). Electronic corpora considerably facilitate search and indexing of documents, and the speed of electronic communication has made it possible to greatly reduce the time to review and publish scholarly work (Spier, 2002). Electronic publishing also permits the distribution of a great many different types of media besides the conventional scholarly article, including datasets, software, videos, and many other forms of supporting material.

The same factors help to facilitate the kind of large-scale peer evaluation described in the present article, of which we already see a great deal of uptake in social networks, video-sharing sites, and other online communities. It is cheap and easy for an individual to rate or comment on a given electronic entity, yet the large-scale of commenting and rating activity enables a great many forms of valuable analysis, that in turn bring benefits back to the evaluating communities (Masum and Zhang, 2004).

One concern related to this approach is that while in principle electronic communication serves to widen access and availability, the practical effect of search, reputation and recommendation tools may in fact be to narrow it (Evans, 2008). On the one hand this may be due to improved filtering of inferior work; however, it is possible that electronic distribution and evaluation systems will heighten the already-known “rich-get-richer” phenomenon

of citation (de Solla Price, 1976; Medo et al., 2011), and perhaps reinforce existing inequalities of attention. One means of addressing this may be to ensure that electronic evaluation systems place a strong focus on diversity as a useful service (Zhou et al., 2010). It certainly emphasizes the point made earlier in this article, that a diversity of metrics is required in order to ensure that the many different types of contribution are all properly recognized and rewarded.

A second concern relates to accessibility. Many of the tools and techniques described here assume ubiquitous access to the internet, something readily available in wealthier nations but still difficult to ensure elsewhere in the world (Best, 2004). Even where access is not an issue, bandwidth may be, for example where the distribution of multimedia files is concerned. However, electronic technologies and communities also serve to *narrow* geographic and economic inequalities, for example making it easier to create documents of equivalent quality (Ginsparg, 1994) and enabling virtual meetings where the cost of travel makes it otherwise difficult for researchers to communicate with their peers (Gichora et al., 2010). The move to online communities as a facilitator of scientific evaluation must certainly be accompanied by a strong push to ensure access.

#### 6.6. OUR VISION FOR FUTURE OF RESEARCH EVALUATION

One of the conclusions that we might draw from the paper is that, as the landscape of the scientific publishing is undoubtedly changing, the processes for the evaluation of research outputs and of researchers are also changing. As we seen in Sections 2, 3, and 4.2, the purpose of the peer review (to find errors or to help improve the paper) is perceived differently by different communities. In the next years we envision the growth of various tools for research evaluation, including open source and those operating with open API/protocols. Such tools would primarily operate on the Web and include the variety of methods for research evaluation, so that PC chairs or journal editors (or even people playing some new emerging roles which do not exist yet) will be able to choose. Examples of tools with such functionalities already emerge (e.g., Mendeley, Peerevaluation.org, Interdisciplines), but it is not yet clear how these tools can be connected and which of them will be adopted widely enough to have a normative effect. We believe that different tools and practices will be adopted by different communities and there is no unique approach that will suit all the researchers on the planet. Moreover, the same researcher working in different contexts will need different tools, and effective evaluation systems should have these choices and alternatives built in by design<sup>14</sup>. With this in mind, attention should be paid less to designing “the” scientific evaluation system of tomorrow – something that, like “the” peer review process, will be an emergent phenomenon based on the different needs of different disciplines and communities. Instead, attention should focus on ensuring interoperability and diversity among the many possible tools that scientific evaluation can make use of.

<sup>14</sup>For instance, Confy, a submission system used by EAI and ICST, allows a choice of various models for conducting peer review – with or without bidding, customizable review forms, and other features. Confy is currently available at <http://cameraready.eai.eu/> and will become open source as the code becomes feature-complete.

## ACKNOWLEDGMENTS

This work has been supported by the EU ICT project LiquidPublication. The LiquidPub project acknowledges the financial support

## REFERENCES

- Adomavicius, G., and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* 17, 734–749.
- Akst, J. (2010). I hate your paper. *Scientist* 24, 36.
- Best, M. L. (2004). Can the internet be a human right? *Hum. Rights Hum. Welf.* 4, 23–31.
- Black, N., van Rooyen, S., Godlee, F., Smith, R., and Evans, S. (1998). What makes a good reviewer and a good review for a general medical journal? *J. Am. Med. Assoc.* 280, 231–233.
- Bornmann, L. (2007). Bias cut. women, it seems, often get a raw deal in science – so how can discrimination be tackled? *Nature* 445, 566.
- Bornmann, L., and Daniel, H.-D. (2005a). Committee peer review at an international research foundation: predictive validity and fairness of selection decisions on post-graduate fellowship applications. *Res. Eval.* 14, 15–20.
- Bornmann, L., and Daniel, H.-D. (2005b). Selection of research fellowship recipients by committee peer review. reliability, fairness and predictive validity of board of trustees' decisions. *Scientometrics* 63, 297–320.
- Bornmann, L., and Daniel, H.-D. (2010a). The validity of staff editors' initial evaluations of manuscripts: a case study of Angewandte Chemie International Edition. *Scientometrics* 85, 681–687.
- Bornmann, L., and Daniel, H.-D. (2010b). The usefulness of peer review for selecting manuscripts for publication: a utility analysis taking as an example a high-impact journal. *PLoS ONE* 5, e11344. doi:10.1371/journal.pone.0011344
- Bornmann, L., Wallon, G., and Ledin, A. (2008). Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European molecular biology organization programmes. *PLoS ONE* 3, e3480. doi:10.1371/journal.pone.0003480
- Brody, T., Harnad, S., and Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *JASIST* 58, 1060–1072.
- Burnham, J. C. (1990). The evolution of editorial peer review. *J. Am. Med. Assoc.* 263, 1323–1329.
- Casati, F., Marchese, M., Mirylenka, K., and Ragone, A. (2010). *Reviewing Peer Review: A Quantitative Analysis of Peer Review*. Technical Report 1813. University of Trento. Available at: <http://eprints.biblio.unitn.it/archive/00001813/>
- Ceci, S. J., and Peters, D. P. (1982). Peer review: a study of reliability. *Change* 14, 44–48.
- Ceci, S. J., and Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3157–3162.
- Cho, M. K., Justice, A. C., Winker, M. A., Berlin, J. A., Waeckerle, J. F., Callahan, M. L., and Rennie, D. (1998). Masking author identity in peer review: what factors influence masking success? PEER Investigators. *JAMA* 280, 243–245.
- de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *J. Am. Soc. Inf. Sci.* 27, 292–306.
- Evans, J. A. (2008). Electronic publication and the narrowing of science and scholarship. *Science* 321, 395–399.
- Fisher, M., Friedman, S. B., and Strauss, B. (1994). The effects of blinding on acceptance of research papers by peer review. *J. Am. Med. Assoc.* 272, 143–146.
- Gichora, N. N., Fatumo, S. A., Ngara, M. V., Chelbat, N., Ramdayal, K., Opat, K. B., Siwo, G. H., Adebisi, M. O., El Gonnouni, A., Zofou, D., Maurady, A. A. M., Adebisi, E. F., de Villiers, E. P., Masiga, D. K., Biz-zaro, J. W., Suravajhala, P., Ommeh, S. C., and Hide, W. (2010). Ten simple rules for organizing a virtual conference – anywhere. *PLoS Comput. Biol.* 6, e1000650. doi:10.1371/journal.pcbi.1000650
- Ginsparg, P. (1994). First steps towards electronic research communication. *Comput. Phys.* 8, 390–396.
- Godlee, F. (2002). Making reviewers visible: openness, accountability, and credit. *JAMA* 287, 2762–2765.
- Godlee, F., Gale, C. R., and Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports a randomized controlled trial. *JAMA* 280, 237–240.
- Goodman, S. N., Berlin, J., Fletcher, S. W., and Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at annals of internal medicine. *Ann. Intern. Med.* 121, 11–21.
- Greaves, S., Scott, J., Clarke, M., Miller, L., Hannay, T., Thomas, A., and Campbell, P. (2006). Overview: Nature's peer review trial. *Nature*. doi: 10.1038/nature05535.
- Harnad, S. (1990). Scholarly skywriting and the prepublication continuum of scientific enquiry. *Psychol. Sci.* 1, 342–344.
- Hu, N., Pavlou, P. A., and Zhang, J. (2009). Overcoming the J-shaped distribution of product reviews. *Commun. ACM* 52, 144–147.
- Ingelfinger, F. J. (1974). Peer review in biomedical publication. *Am. J. Med.* 56, 686–692.
- Jefferson, T., Rudin, M., Folse, S. B., and Davidoff, F. (2007). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane* 41, MR000016.
- Jefferson, T., Wager, E., and Davidoff, F. (2002a). Measuring the quality of editorial peer review. *JAMA* 287, 2786–2790.
- Jefferson, T., Alderson, P., Wager, E., and Davidoff, F. (2002b). Effects of editorial peer review: a systematic review. *JAMA* 287, 2784–2786.
- Justice, A. C., Cho, M. K., Winker, M. A., Berlin, J. A., Rennie, D., and PEER Investigators. (1998). Does masking author identity improve peer review quality? A randomized controlled trial. *JAMA* 280, 240–242.
- Kassirer, J. P., and Campion, E. W. (1994). Peer review: crude and understudied, but indispensable. *J. Am. Med. Assoc.* 272, 96–97.
- Katz, D. S., Proto, A. V., and Olmsted, W. W. (2002). Incidence and nature of unblinding by authors: our experience at two radiology journals with double-blinded peer review policies. *Am. J. Roentgenol.* 179, 1415–1417.
- Kronick, D. A. (1990). Peer review in 18th-century scientific journalism. *JAMA* 263, 1321–1322.
- Lawrence, P. A. (2003). The politics of publication. *Nature* 422, 259–261.
- Lee, K., Boyd, E., Holroyd-Leduc, J., Bacchetti, P., and Bero, L. (2006). Predictors of publication: characteristics of submitted manuscripts associated with acceptance at major biomedical journals. *Med. J. Aust.* 184, 621.
- Link, A. M. (1998). Us and non-US submissions: an analysis of reviewer bias. *JAMA* 280, 246–247.
- Lock, S. (1994). Does editorial peer review work? *Ann. Intern. Med.* 121, 60–61.
- Lynch, J. R., Cunningham, M. R., Warne, W. J., Schaad, D. C., Wolf, F. M., and Leopold, S. S. (2007). Commercially funded and united states-based research is more likely to be published; good-quality studies with negative outcomes are not. *J. Bone Joint Surg. Am.* 89, 1010–1018.
- Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H.-D., and O'Mara, A. (2009). Gender effects in the peer reviews of grant proposals: a comprehensive meta-analysis comparing traditional and multilevel approaches. *Rev. Educ. Res.* 79, 1290–1326.
- Masum, H., and Zhang, Y.-C. (2004). Manifesto for the reputation society. *First Monday* 9 [Online].
- McCook, A. (2006). Is peer review broken? *Scientist* 20, 26.
- McNutt, R. A., Evans, A. T., Fletcher, R. H., and Fletcher, S. W. (1990). The effects of blinding on the quality of peer review: a randomized trial. *JAMA* 263, 1371–1376.
- Medo, M., Cimini, G., and Gualdi, S. (2011). Temporal effects in the growth of networks. Available at: <http://arxiv.org/abs/1109.5560>
- Medo, M., and Wakeling, J. R. (2010). The effect of discrete vs. continuous-valued ratings on reputation and ranking systems. *Europhys. Lett.* 91, 48004.
- Odlyzko, A. M. (1995). Tragic loss or good riddance? The impending demise of traditional scholarly journals. *Int. J. Hum. Comput. Sci.* 42, 71–122.
- Olson, C. M., Rennie, D., Cook, D., Dickersin, K., Flanagan, A., Hogan, J. W., Zhu, Q., Reiling, J., and Pace, B. (2002). Publication bias in editorial decision making. *JAMA* 287, 2825–2828.
- Ophof, T., Coronel, R., and Janse, M. J. (2002). The significance of the peer review process against the background of bias: priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. *Cardiovasc. Res.* 56, 339–346.
- Osman, N., Sabater-Mir, J., and Sierra, C. (2011). "Simulating research behaviour," in *12th International Workshop on Multi-Agent-Based Simulation (MABS'11)*, Taipei.

- Osman, N., Sabater-Mir, J., Sierra, C., de Pinninck Bas, A. P., Imran, M., Marchese, M., and Ragone, A. (2010a). *Credit attribution for liquid publications*. Deliverable D4.1, Liquid Publications Project. Available at: [https://dev.liquidpub.org/svn/liquidpub/papers/deliverables/LP\\_D4.1.pdf](https://dev.liquidpub.org/svn/liquidpub/papers/deliverables/LP_D4.1.pdf)
- Osman, N., Sierra, C., and Sabater-Mir, J. (2010b). "Propagation of opinions in structural graphs," in *ECAI 2010: Proceedings of the 19th European Conference on Artificial Intelligence, Vol. 215 of Frontiers in Artificial Intelligence and Applications*, eds H. Coelho, R. Studer, and M. Wooldridge (Lisbon: IOS Press), 595–600.
- Parra, C., Birukou, A., Casati, F., Saint-Paul, R., Wakeling, J. R., and Chlamtac, I. (2011). "UCount: a community-driven approach for measuring scientific reputation," in *Proceedings of Altimetrics11: Tracking Scholarly Impact on the Social Web*, Koblenz.
- Purcell, G. P., Donovan, S. L., and Davidoff, F. (1998). Changes to manuscripts during the editorial process: characterizing the evolution of a clinical paper. *J. Am. Med. Assoc.* 280, 227–228.
- Ragone, A., Mirylenka, K., Casati, F., and Marchese, M. (2011). "A quantitative analysis of peer review," in *13th International Society of Scientometrics and Informetrics Conference*, Durban.
- Reinhart, M. (2009). Peer review of grant applications in biology and medicine: reliability, fairness, and validity. *Scientometrics* 81, 789–809.
- Ross, J. S., Gross, C. P., Desai, M. M., Hong, Y., Grant, A. O., Daniels, S. R., Hachinski, V. C., Gibbons, R. J., Gardner, T. J., and Krumholz, H. M. (2006). Effect of blinded peer review on abstract acceptance. *J. Am. Med. Assoc.* 295, 1675–1680.
- Sense About Science. (2009). *Peer Review Survey: Preliminary Results*. Available at: <http://www.senseaboutscience.org.uk/index.php/site/project/29/>
- Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *J. R. Soc. Med.* 99, 178–182.
- Spier, R. (2002). The history of the peer-review process. *Trends Biotechnol.* 20, 357–358.
- Swan, A. (2007). Open access and the progress of science. *Am. Sci.* 95, 198–200.
- Underwood, A. J. (2004). It would be better to create and maintain quality rather than worrying about its measurement. *Mar. Ecol. Prog. Ser.* 270, 283–286.
- van Rooyen, S., Black, N., and Godlee, F. (1999). Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. *J. Clin. Epidemiol.* 52, 625–629.
- Walsh, E., Rooney, M., Appleby, L., and Wilkinson, G. (2000). Open peer review: a randomised controlled trial. *Br. J. Psychiatry* 176, 47–51.
- Ware, M., and Monkman, M. (2008). *Peer Review in Scholarly Journals: Perspective of the Scholarly Community – An International Study*. Survey Commissioned by the Publishing Research Consortium. Available at: <http://www.publishingresearch.net/PeerReview.htm>
- Wenneras, C., and Wold, A. (1997). Nepotism and sexism in peer-review. *Nature* 387, 341–343.
- Zhou, T., Kuscsik, Z., Liu, J.-G., Medo, M., Wakeling, J. R., and Zhang, Y.-C. (2010). Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4511–4515.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 July 2011; paper pending published: 07 August 2011; accepted: 11 November 2011; published online: 14 December 2011.

Citation: Birukou A, Wakeling JR, Bartolini C, Casati F, Marchese M, Mirylenka K, Osman N, Ragone A, Sierra C and Wassef A (2011) Alternatives to peer review: novel approaches for research evaluation. *Front. Comput. Neurosci.* 5:56. doi: 10.3389/fncom.2011.00056  
Copyright © 2011 Birukou, Wakeling, Bartolini, Casati, Marchese, Mirylenka, Osman, Ragone, Sierra and Wassef. This is an open-access article subject to a non-exclusive license between the authors and Frontiers Media SA, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and other Frontiers conditions are complied with.



# Fair and open evaluation may call for temporarily hidden authorship, caution when counting the votes, and transparency of the full pre-publication procedure

Talis Bachmann\*

Department of Cognitive and Forensic Psychology, Institute of Public Law, University of Tartu, Tartu, Estonia

\*Correspondence: talis.bachmann@ut.ee

## THE PROBLEM OF IMBALANCE

The traditional system of manuscript evaluation has created a certain imbalance bordering with unfairness: while the authors of submitted papers typically have had their identity disclosed already at the outset, the reviewers have remained mostly anonymous. With a new open evaluation system being currently envisaged, the main difference would be that evaluators become disclosed as well, which is a significant step toward balance and fairness. More openness and constructive interactivity in the reviewing process have become to be practiced increasingly more, including some noteworthy success (e.g., the Shepherding system at the European Conference on Pattern Languages of Programming and Computing and the Frontiers initiative). However, while both the authors and the reviewers have become disclosed, the past collective “instincts” and traditions of reviewers as evaluators could often remain unaffected by this change. Thus, if the new envisaged evaluation system fully discloses both the authors and evaluators, it nevertheless cannot tackle all potential sources of bias and unfairness. Although for the majority of researchers this system seems to be suitable, there are also researchers who feel that not all possible sources of unfairness would be eliminated. Therefore there should be also an additional optional format of review and publishing that goes even further in pursuit for minimizing the impact of subjectivity. Why so?

## THE SOURCES OF UNFAIRNESS

First, *de facto* scientific policies have always featured certain elements of paradigmatic power-structure, impact of authority, regional interests, “pecking order.” This does not necessarily constitute a bias or animosity toward particular people, but

rather a negative bias against “alien” theoretical approaches and positive attitudes in adhering to traditional views or views of the most authoritative scientists. Second, the history of personal relationships between authors (and/or their colleagues) on the one hand and evaluators (and/or their colleagues) on the other hand, may prejudice the whole process. (This includes an earnestly perceived but non-deliberately distorted understanding of the papers and views.) Third, as some field or tradition of research may be willy-nilly in a stage of stagnation, new ideas and approaches can be almost collectively resisted and negatively evaluated. Therefore, it is advisable to adopt two additional, even if not universally implemented, formats of evaluation of the written work. (1) Keeping the identity of author(s) undisclosed for up to a year post-publication (with later disclosure) if the author(s) wish so. This should diminish the author’s identity-based negative biases. (2) As science is inherently paradigmatic and because a large number of evaluators are inevitably accustomed to the currently prevailing paradigms, weighing votes or numbers of positively or negatively valenced evaluations can often be biased toward reactionary or conventional views. This weighing style of evaluation may also result in an opposite bias of (sometimes) undeserved praise and highly positive rating of dull or non-innovative works deriving from scientific-political influences and habits. Both of these biases should also be counterbalanced in the new envisaged evaluation system.

## POSSIBLE REMEDIES AND DESIGN

In order to alleviate the above-mentioned problems, in case of each paper submission and the weighing procedure the following principles could be adopted. A set of concomitant open writings of evalua-

tion are published in the finalized issue of the periodical together with the main article. Similarly to what has been practiced by Behavioral and Brain Sciences (BBS, Cambridge University Press) these evaluation papers may be accompanied by the authors’ reply and counter-criticism. Furthermore, the relatively informal pre-publication stage of preparing a paper and its comments should be transparent – all interested and involved parties can access all the submitted main-article manuscripts as well as all review/evaluation papers. In other words, the full portfolio of submissions by professional authors and full set of reviews should be transparent and made available for the scientific community. (The currently available electronic means help to overcome the endangering capacity problems.) Seeds of this format have been planted already by such outlets as BBS, Interdisciplines (supported by OpinioNet and LiquidPub), and some others. The open review could also adapt the format suggested by Lee (2011) in his Selected-Papers Network model: reviewers can endorse a paper for publication and also publish a concomitant review. After some critical time has elapsed, the unpublished submissions and reviews will eventually be removed from the public domain if authors wish so, but may also remain accessible under the label “unpublished.” The original timeframe with full disclosure has made it possible to copy the pre-publication versions of main papers and critical evaluative papers by all professionals interested anyway. It should be allowed, where necessary, to cite also the unpublished but temporarily accessible “pre-publication” works and data included there. How could this vision relate to the central design decisions when constructing a new system for open evaluation? The backbone of the procedure could look something like this:



pre-acceptance screening > **open review** > (non)acceptance > **publication/closure** > **post-review**.

*Pre-acceptance screening.* Some minimum screening for the obviously non-professional or mocking contributions or technically/formatively clearly non-conforming works is applied. This is a non-transparent step 1, based on editorial decision.

*Reviewing.* Step 2 marks the beginning of a review process, which in turn means a fully transparent display of both the complete submitted material as well as a full set of comments by reviewers and editors. As for the alternative metrics (e.g., paper downloads), I suggest not using this as a standard procedure in the reviewing stage. It would burden the already voluminous body of text in the evaluation treatment; furthermore, downloads are heavily biased by non-substantial factors such as journal rankings, visibility, and influence of authors, etc. Downloads data could be made accessible at request, not attached/displayed by default.

Therefore, it is important to guarantee that scientific objectivity prevails and political motivations are minimized. (i) Papers become published after minimal review, more thorough post-review and criticism follows publication. Criticism is highly professional and well-informed allowing for substantive commentary elements just as an old Estonian proverb suggests – the wolves are fed and the sheep alive. (ii) Special explicit sections or footnotes in the form of a short commentary regarding the views and theories that be in question, why so, and with what implications are advisable. (iii) The system should resort to transparent signed reviews and ratings. On the other hand, about 1/4 of future open-access journals could remain “traditional” in terms of anonymity of reviewers if they wish so. The authors can choose the type of journal they wish to be published in. (iv) Evaluation *may* continue in the post-publication review phase and for a considerable length of time (e.g., with promising or controversial papers, papers with possibly controversial or limited results), but need not. (v) Ratings should be used only if differentiated and specific enough – e.g.,

novelty of interpretations/theory, technical quality, methodological advances, discovery status, creativity of ideas, etc. Ratings should not be automatically revealed together with a paper, but only accessible at request by readers.

(Non)acceptance is step 3 followed by *publication* or closure (step 4). Published papers get their final unique article identification label with specification of volume/issue/pages/web-link/date added to the initial identifier attributed to the manuscript at submission. Unpublished papers keep their unique initial identifier supplemented by the label “closed.”

*Post-publication affairs.* This stage is optional, depending on evoked interest, potential reviewers’ incentives, new emerging circumstances, etc. In the post-publication evaluative open review (step 5) by the original or new reviewers, formatted as separate brief commentaries, the emphasis in informed comments would expectedly shift more toward refined debates, which remains an open discussion forum for quite long time (unless it dies out). The continuing evaluation should be useful because not everybody who may have seen the paper earlier and because some important evidence and related new results may appear just a bit later. On the other hand, evaluative prioritization and rankings based on downloads statistics etc., allowing readers to compare different papers should be only accessible at the readers’ request, but not publicly displayed. The time period covering months and a couple of years post-publication is too short for real evaluation that would stand the test of time, scientific-political factors and underdevelopment of the field of research may interfere too much with substance, and there are too many reasons for downloads other than that a paper is of really high quality, important, or truly innovative. It is questionable to evaluate the value of a scientific publication by numbers of downloads precisely for the above reasons. Let the citation databases live their separate lives without mixing publishing business with scientome-

trics. The upon-request post-publication ratings should be differentiated and concrete rather than based on overall general statistics. When considering whether to adopt comprehensive rules or varying formats for defining the evaluation formulae, we should leave some options open. Although publishers (i.e., collectives of scientists) may try to reach a consensus in unification, some other optional instrument should also be developed, e.g., authors may be allowed to evaluate their contribution in terms of ratings along various evaluation scales.

#### General strategies and specific formats.

However, there should be a special publication format optional for use and even recommended to the authors, i.e., to remain anonymous for a year post-publication. The articles are cited for this period authored as temporarily anonymous (*anon-temp*). As soon as the year has passed authorship disclosure becomes compulsory. Each article, whether in the anonymity stage or post-disclosure stage, has a unique identifier which helps to be certain that the same article is referred to. (It is widely believed that despite attempts to remain anonymous, professional readers can in practice successfully guess the author’s identity. Preliminary information available from conferences, lab visits, previous publications, etc., could make it doubtful to guarantee anonymity. Anonymity also may discourage researchers from taking credit for their achievements and fostering one’s career. All this can be countered by special care in writing an article and optimizing the frequency of opting for one or another type of publication.) Most importantly, this new format of publication may not become a prevailing option, but an outlet especially useful for innovative research and cases where authors feel the need to remain anonymous for the time being and therefore take care in not including disclosing information in their papers.

## ASPECTS OF IMPLEMENTATION

How can we efficiently bring about a transition toward the future system? There will be inevitably some period of trial-and-

error and perhaps the development should continue even further. However, there are some threats that the new system may not turn out as was originally expected or it could make a mockery out of what was initially envisaged as an aspiration toward fairness, speed, and openness. There tend to be two kinds of scholars – researchers immersed in high-quality top-level research vs scientific administrators and organizers, not so prominent as scientists, but influential in other ways. The former are not eager to devote time to implementing reforms and organizational matters whereas for the latter, reforms and “the so-called reforms” are their natural domain. Consequently, the future evaluation system may not attract many truly informed and complex-free academics as evaluators, but too many fresh post-docs and scientific administrators instead and thus the new system may fail to achieve its goals. An idealist hope is characterized by the following: new open-access periodicals will be managed and the tone set by teams of top-level scientists who are known for their objectivity, generosity, sharp analytical vision, love of creativity, and innovation, with preference for substance rather than nice “packaging,” and possession of wide contextuated knowledge combined with the ability to create new knowledge instead of the mere familiarity with the currently prevailing buzzwords. This group of scientists-by-heart will invite the best papers and the reform will be implemented through “magnetism” toward the highest impact, fast-track publication outlets. Furthermore, can the new journals, in minority among the prevailing earlier system survive the already established environment? It is hard to know and only time would tell whether the actual demand for this format of publishing will help its survival.

There is yet another threat. Underdeveloped countries with less financial and scientific-political power will have fewer chances to publish and wield influence as their institutions simply have a limited budget. The promise to take this into account is just an excuse and cannot be applied endlessly for financial reasons. Moreover, abandoning publication fees altogether would be embarrassing for the authors or their institution. So, a “promise of discrimination” is lurking behind the open-access, pay-per-publication system. It would typically result in a situation where in order to overcome this obstacle, the less prosperous researchers will “sell” their ideas for joint authorship and although it might entail an essentially positive aspect of international integration and co-operation this also means that their scientific production will be controlled from outside of their own environment. As a remedy, I suggest the possibility of dispersing the leading open-access journals’ teams and facilities geographically in terms of choices/appointments of editors, editorial board members and reviewers, IT-facilities servicing a journal, etc. It is also important to avoid the excess of reviewer monopoly such as about three to eight authors having recently been published in a particular paradigm, review most of the submissions anonymously, including the review of their direct competitors. The excess reviewing by currently visibly publishing post-docs should be also moderated because many of them often tend to have too narrow a perspective, knowledge and expertise related strictly to their PhD topic without a broader contextuated knowledge and experience. (This is despite the fact that they tend to be more absorbed by the reviewing process and may be even better in spotting the errors. However, according to my own extended experience with younger reviewers and fresh researchers, they tend to lack multifaceted, broad view and sufficient

knowledge of the host of earlier published research.) The scope of reading even by several different reviewers may lack sufficient depth. Now, this is precisely the place where a fully transparent pre-publication evaluation system together with the continuing post-publication discussion may have its advantages over the traditional system.

In *conclusion*, the key proposals introduced above contain the following: pre-publication manuscripts selected for review and the reviewer’s work are both transparent, the reviewer’s identity is disclosed; the author of a paper may remain anonymous; discussion of a paper can continue post-publication; overly critical or overly flattering evaluation can be at least minimally counterbalanced; the author has an option to remain temporarily anonymous post-publication. Measures should be taken against bureaucratizing and politicizing the new review system, the choice of reviewers should not be restricted to junior scientists or “activists,” the new system of review and publishing should be introduced also in the less-developed regions accompanied by lower pay-per-publication costs. Last but not least, the traditional system of journal publishing should not be discarded instead it should be preserved as a viable option.

## REFERENCE

Lee, C. (2011). Open peer review by a Selected-Papers Network. *Front. Comput. Neurosci.* (in press).

Received: 09 January 2011; accepted: 11 December 2011; published online: 29 December 2011.

Citation: Bachmann T (2011) Fair and open evaluation may call for temporarily hidden authorship, caution when counting the votes, and transparency of the full pre-publication procedure. *Front. Comput. Neurosci.* 5:61. doi: 10.3389/fncom.2011.00061

Copyright © 2011 Bachmann. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Open peer review by a selected-papers network

Christopher Lee<sup>1,2\*</sup>

<sup>1</sup> Department of Chemistry and Biochemistry, UCLA-DOE Institute for Genomics and Proteomics, University of California Los Angeles, Los Angeles, CA, USA

<sup>2</sup> Department of Computer Science, UCLA-DOE Institute for Genomics and Proteomics, University of California Los Angeles, Los Angeles, CA, USA

## Edited by:

Diana Deca, University of  
Amsterdam, Netherlands

## Reviewed by:

Diana Deca, University of  
Amsterdam, Netherlands  
Rogier Kievit, University of  
Amsterdam, Netherlands  
Aliaksandr Birukou, European Alliance  
for Innovation, Italy  
Dwight Kravitz, National Institutes of  
Health, USA

## \*Correspondence:

Christopher Lee, Department of  
Chemistry and Biochemistry,  
University of California Los Angeles,  
Los Angeles, CA 90095, USA.  
e-mail: leec@chem.ucla.edu

A selected-papers (SP) network is a network in which researchers who read, write, and review articles subscribe to each other based on common interests. Instead of reviewing a manuscript in secret for the Editor of a journal, each reviewer simply publishes his review (typically of a paper he wishes to recommend) to his SP network subscribers. Once the SP network reviewers complete their review decisions, the authors can invite any journal editor they want to consider these reviews and initial audience size, and make a publication decision. Since all impact assessment, reviews, and revisions are complete, this decision process should be short. I show how the SP network can provide a new way of measuring impact, catalyze the emergence of new subfields, and accelerate discovery in existing fields, by providing each reader a *fine-grained filter for high-impact*. I present a three phase plan for building a basic SP network, and making it an effective peer review platform that can be used by journals, conferences, users of repositories such as arXiv, and users of search engines such as PubMed. I show how the SP network can greatly improve review and dissemination of research articles in areas that are not well-supported by existing journals. Finally, I illustrate how the SP network concept can work well with existing publication services such as journals, conferences, arXiv, PubMed, and online citation management sites.

**Keywords:** open evaluation, peer review, scientometrics, journal, publishing, interdisciplinary research

## 1. INTRODUCTION

### 1.1. GOALS: WHAT PROBLEMS DOES THIS PROPOSAL AIM TO SOLVE?

I begin by briefly outlining the problems in existing peer review that this proposal aims to resolve. Here I only *define* the problem, to motivate the subsequent proposal. I will also briefly state some issues that I explicitly *exclude* from its goals, to make my focus clear.

Current peer review suffers from systemic blind spots, bottlenecks, and inefficiencies that retard the advance of research in many areas. These pathologies reflect the petrification of peer review from what it started as (informal discussions of a colleague's latest report in a club meeting) into a rigid system of assumptions inherited from outdated distribution and communication models (ink-on-paper printing press and postal mail). Peer review started out as a PULL model (i.e., each person decides what to receive – concretely, which talks to attend), but petrified into a PUSH model (i.e., a centralized distribution system decides what everyone else should receive). Most of these pathologies are due to the basic mismatch of the PUSH model versus the highly specialized, interdisciplinary, and rapidly evolving nature of scientific research. This proposal seeks to address the following problems:

- *Expert peer review (EPR) does not work for interdisciplinary peer review (IDPR).* EPR means the assumption that the reviewer is expert in all aspects of the paper, and thus can evaluate *both* its impact and validity, and can evaluate the paper prior to obtaining answers from the authors or other referees. IDPR means the situation where at least one part of the paper lies outside the reviewer's expertise. Since journals universally assume EPR, this creates artificially high barriers to innovative papers

that combine two fields (Lee, 2006) – one of the most valuable sources of new discoveries.

- *Shoot first and ask questions later* means the reviewer is expected to state a REJECT/ACCEPT position *before* getting answers from the authors or other referees on questions that lie outside the reviewer's expertise.
- *No synthesis:* if review of a paper requires synthesis – combining the different expertise of the authors and reviewers in order to determine what assumptions and criteria are valid for evaluating it – both of the previous assumptions can fail badly (Lee, 2006).
- *Journals provide no tools for finding the right audience for an innovative paper.* A paper that introduces a new combination of fields or ideas has an *audience search* problem: it must search multiple fields for people who can appreciate that new combination. Whereas a journal is like a TV channel (a large, pre-defined audience for a standard topic), such a paper needs something more like Google – a way of quickly searching multiple audiences to find the subset of people who can understand its value.
- *Each paper's impact is pre-determined rather than post-evaluated:* By “pre-determination” I mean that both its impact metric (which for most purposes is simply the title of the journal it was published in) and its actual readership are locked in (by the referees' decision to publish it in a given journal) before any readers are allowed to see it. By “post-evaluation” I mean that impact should simply be measured by the research community's long-term response and evaluation of it.
- *Non-expert PUSH* means that a pre-determination decision is made by someone *outside* the paper's actual audience, i.e.,

the reviewer would not ordinarily *choose* to read it, because it does not seem to contribute sufficiently to his personal research interests. Such a reviewer is forced to guess whether (and how much) the paper will interest *other* audiences that lie outside his personal interests and expertise. Unfortunately, people are not good at making such guesses; history is littered with examples of rejected papers and grants that later turned out to be of great interest to many researchers. The highly specialized character of scientific research, and the rapid emergence of new subfields, make this a big problem.

In addition to such *false-negatives*, non-expert PUSH also causes a huge false-positive problem, i.e., reviewers accept many papers that do not personally interest them and which turn out not to interest anybody; a large fraction of published papers subsequently receive zero or only one citation (even including self-citations; Adler et al., 2008). Note that non-expert PUSH will occur by default unless reviewers are instructed to *refuse* to review anything that is not of compelling interest for *their own work*. Unfortunately journals assert an opposite policy.

- *One man, one nuke* means the standard in which a single negative review equals REJECT. Whereas post-evaluation measures a paper's value over the whole research community ("one man, one vote"), standard peer review enforces conformity: if *one* referee does not understand or like it, prevent *everyone* from seeing it.
- *PUSH makes refereeing a political minefield*: consider the contrast between a conference (where researchers publicly speak up to ask challenging questions or to criticize) vs. journal peer review (where it is reckoned necessary to hide their identities in a "referee protection program"). The problem is that each referee is given artificial power over what other people can like – he can either confer a large value on the paper (by giving it the imprimatur and readership of the journal) or consign it zero value (by preventing those readers from seeing it). This artificial power warps many aspects of the review process; even the "solution" to this problem – shrouding the referees in secrecy – causes many pathologies. Fundamentally, current peer review treats the reviewer not as a *peer* but as one who wields a *diktat*: prosecutor, jury, and executioner all rolled into one.
- *Restart at zero* means each journal conducts a completely separate review process of a paper, multiplying the costs (in time and effort) for publishing it in proportion to the number of journals it must be submitted to. Note that this particularly impedes innovative papers, which tend to aim for higher-profile journals, and are more likely to suffer from referees' IDPR errors. When the time cost for *publishing* such work exceeds by several fold the time required to *do* the work, it becomes more cost-effective to simply abandon that effort, and switch to a "standard" research topic where repetition of a pattern in many papers has established a clear template for a publishable unit (i.e., a widely agreed checklist of criteria for a paper to be accepted).
- *The reviews are thrown away*: after all the work invested in obtaining reviews, no readers are permitted to see them. Important concerns and contributions are thus denied to the research community, and the referees receive no credit for the vital contribution they have made to validating the paper.

In summary, current peer review is designed to work for large, well-established fields, i.e., where you can easily find a journal with a high probability that every one of your reviewers will be in your paper's target audience and will be expert in all aspects of your paper. Unfortunately, this is just not the case for a large fraction of researchers, due to the high level of specialization in science, the rapid emergence of new subfields, and the high value of boundary-crossing research (e.g., bioinformatics, which intersects biology, computer science, and math).

I wish to list explicitly some things that this proposal does *not* seek to change:

- it does not seek to *replace* conventional journals but rather to *complement* them by offering an improved peer review process.
- it does not seek to address large audience distribution channels (e.g., marquee journals like *Nature*, or journals associated with large, well-established fields), or papers that fit these journals well. Instead it focuses on papers that need to actively *search* for an audience, e.g., because they are at the intersection of multiple audiences.
- it does not address the large fraction of papers published by journals that do not interest anyone (as indicated by lack of subsequent citation). Instead it focuses on papers for which it can find an audience that considers the paper "must-read."

## 2. THE PROPOSAL IN BRIEF

### 2.1. WHAT IS A SELECTED-PAPERS NETWORK?

Here I briefly summarize the proposal, by sketching its system for peer review. My purpose is to define the proposed system clearly, and to highlight its core principles. Note that this section will neither seek to prove that it solves all the problems above, nor address the political question of how to make the current system yield to the proposed system. Those are separate issues that deserve separate treatment. Core principles:

- Instead of reviewing a manuscript in secret for the Editor of a journal, a referee simply publishes his review (typically of a paper he wishes to recommend) on an open *Selected-Papers (SP) network*, which automatically forward his review to readers who have subscribed to his *selected-papers list* because they feel his interests match their own, and trust his judgment. I will refer to such a reviewer as a "selected-paper reviewer" (SPR).
- Instead of submitting a paper to a specific journal, authors submit it to the SP network, which quickly scans a large number of possible reviewers to see if there is an audience that considers it "must-read." This audience search process should take just a few days using automated e-mail and click-through metrics. This determination is direct: the system simply measures whether seeing the title makes someone click to see the abstract; whether seeing the abstract makes them click to see the text; whether seeing the text makes them click to see the figures etc.
- Reviewers are instructed to only consider papers that are of compelling interest for their own work, i.e., that they would eagerly choose to read even if they were not being asked to review. In other words, each reviewer should represent only his own interests, and should not try to guess whether it will interest other audiences. Following this principle, refusing to review a paper



is itself a review (“this paper does not interest me enough to read”). During this pre-review phase, each SPR can informally ask questions or make comments without yet committing to review the paper, and can restrict their comments to be visible to the authors only, the other reviewers as well, or as part of the permanent review record for the paper that will become public if the paper is published.

- If no one considers the paper must-read and is willing to review it, no further action is needed. (The authors can send it to a regular journal if they wish).
- Otherwise, the SPRs who agree to act as reviewers begin a Questions/Answers phase where they raise whatever questions or issues they want, to assess the validity of the paper. A reviewer can opt to remain anonymous if he feels this is necessary. The authors and referees work together to identify and resolve these issues in the context of an issue tracking system like those used for debugging a software release. This phase would have a set deadline (e.g., 2 weeks). If the authors undertake a major revision (e.g., with new data), a new 2 week Q/A phase ensues.
- Next, during the assessment phase the reviewers individually negotiate with the authors over validation issues they consider essential, e.g., “If you do this additional control, that would address my concern and I could recommend your paper.”
- The authors decide how much they are willing to do for the *final version* of the paper, based on their time pressures and other competing interests. They produce this final version.
- Each reviewer decides whether or not to recommend the final paper to their subscribers. This gives the paper a known initial audience size (the total number of subscribers of the reviewers who choose to recommend it).
- Once the reviews are complete, they can now be considered by a journal or conference editor. The authors invite any editor they want to consider these reviews and initial audience size, and make a publication decision. Since all impact assessment, reviews, and revisions are complete, this process should be short, e.g., the editor should be given a deadline of a week or so to reach a decision. Note that since many reviewers are also editors, the reviewers’ decisions may already confer a guaranteed publication option. For example, for many fields there is a high probability that one of the reviewers would be a *PLoS ONE* Academic Editor and thus could unilaterally decide to accept the paper to *PLoS ONE*. Note that the journal *should not* seek to re-review the paper using their own procedures or ask the original reviewers to give them a new decision (“is this good enough for *Nature*?”). This is a clean division of labor: the reviewers decide impact *for themselves* (and no one else) and assess validity; the journal decides whether the paper is appropriate for the journal’s audience.
- The journal publishing the paper may ask the authors to reformat it, but should not alter the content of the final version (it might be acceptable to have some sections published online but not in the print version).
- When the paper is published online, the reviewers’ recommendations of the paper are forwarded to their subscribers, with a link to view the paper wherever it is published (e.g., the journal’s website). Thus the journal benefits from not only the free

review process, but also the free targeted marketing of the paper provided by the SP network.

- The reviews themselves would be published online. Positive recommendations could be published in a “News and Views” style journal created for this purpose; negative reviewers could opt to publish a brief “Letters” style critique in a journal created for this purpose (“Critical Reviews in. . .”). The community should have access to these important validation assessments, and reviewers should receive credit for this important contribution. All review comments and issues would be available in the SP network page for the paper, which would remain open as the forum for long-term evaluation of the paper’s claims. That is, other users could raise new issues or report data that resolve issues.
- Any online display of the paper’s title, abstract, or full content (e.g., on PubMed, or the journal’s website) should include recommendation links showing who recommended it, each linked to a page showing the text of the review, and that reviewer’s other recommendations/reviews. This would enable readers who find a paper they like to find reviewers who share their interests, and subscribe to receive their future recommendations.
- Furthermore, each reader who considers the paper must-read should add it to their own recommendation list (which at this point does not require writing a review, since the paper is already published). They would simply click a “Like!” icon on any page displaying the paper, with options to simply cite the paper, or recommend it to their own subscribers. In this way the paper can spread far beyond its initial audience – but only if new readers continue to find it “must-read.” This constitutes true impact measurement via post-evaluation: each person decides for themselves what the paper’s value is to them, and the system reports this composite measurement over all audiences.

## 2.2. IMMEDIATE PAYOFFS AND REDUCED BARRIERS TO ENTRY

Systemic reform always faces a bootstrap problem: early adopters gain little benefit (because no one else is participating in the new system yet) and suffer high costs. I have designed this proposal specifically to solve this bootstrap problem by giving it immediate payoffs for the key players (referees, readers, journals, and authors) and to allow it to begin working immediately *within* the existing system.

- For *reviewers*, there would normally be little incentive to review manuscripts in a new system, because doing so would have little impact (initially they would have no subscribers). The SP network solves this in two ways: first, by simply making itself a peer review platform for submission to existing journals (so the reviewer has just as much impact and incentive as when they review for an existing journal); second, by displaying their paper recommendations on the key sites where readers find papers (e.g., PubMed, journal websites etc.). This would give their recommendations a large audience even before they have any subscribers, and would create a fast path for them to gain subscribers.
- *Readers* would ordinarily have little incentive to join a new subscription system, if it requires them to change how they find papers (e.g., by having to log in to a new website). After all, there will initially be very few reviewers or recommendations



in the system, and therefore little benefit for readers. The SP network solves this by displaying its recommendations within the main websites where readers find papers, e.g., PubMed and journal websites. Readers will see these recommendations even if they are not subscribers, and if they find them valuable, will be able to subscribe with a single click.

- *Journals* may well look askance at any proposal for change. However, this proposal offers journals immediate benefits while preserving their autonomy and business model. On the one hand, the SP network provides free marketing for the journal's papers, in the form of *recommendations* that will send traffic to the journal, and *subscribers* who provide a guaranteed initial readership for a recommended paper. What journal would *not* want recommendations of its papers to be shown prominently on PubMed and its own website? On the other hand, the SP network will cut the journal's costs by providing it with free reviewing services that go far beyond what journals do, e.g., active *audience search* and direct measurement of impact over multiple audiences. Since reviewers will still have the option to review anonymously, the journal cannot claim the process is less rigorous (actually, it will be more rigorous, due to its greatly improved discussion, and sharing of multiple expertises). Moreover, the journal preserves complete autonomy over both its editorial decision-making and its business model. It seems reasonable to expect that multiple journals (e.g., the PLoS family) would quickly agree to become SP network partners, i.e., they would accept paper submissions via the SP network review process.
- *Authors* would ordinarily have little incentive to send their papers to a new subscription system rather than to an existing journal. After all, initially such a system will have few subscribers, and no reputation. The SP network solves this by acting as a peer review platform for submitting a paper to existing journals. Indeed, it offers authors a signal advantage over directly submitting to a journal: a unified review process that guarantees a single round of review; i.e., even if the paper is rejected by one or more journals, it will *not* need to be re-reviewed. This is a crucial advantage, e.g., for papers whose validity is solid but where the authors want to "gamble" on trying to get it into a high-impact journal.

### 3. BENEFITS OF A SELECTED-PAPERS NETWORK

#### 3.1. BENEFITS FOR READERS

The core logic of the SP network idea flows from inherent inefficiencies in the existing system.

*For readers, journals no longer represent an efficient way to find papers that match their specific interests.* In paper-and-ink publishing, the only way to make distribution cost-effective was to rely on economies of scale, in which each journal must have a large audience of subscribers, and delivers to every subscriber a uniform list of papers that are supposedly all of interest to them. In reality, most papers in any given journal are simply not of direct interest to (i.e., specifically relevant to the work of) each reader. For example, in my own field the journal *Bioinformatics* publishes a very large number and variety of papers. The probability that any one of these papers is of real interest to my work is low. For this reason, readers no longer find papers predominantly by "reading a journal"

from beginning to end (or even just its table of contents). Instead, they have shifted to finding papers mainly from literature searches (PubMed, National Library of Medicine, 1996; Google, Google Scholar, Acharya and Verstak, 2005; etc.) and word of mouth. Note that the latter is just an informal "Selected-Papers network."

For readers, an SP network offers the following compelling advantages:

- *Higher relevance.* Instead of dividing attention between a number of journals, each of which publishes only a small fraction of directly relevant papers, a reader subscribes (for free) to the Selected-Papers lists of peers whose work matches his interests, and whose judgment he trusts. Note that since most researchers have multiple interests, you typically subscribe specifically to just the recommendations from a given SPR that are in *your* defined areas of interest. The advantage is fundamental: whereas journals lump together papers from many divergent subfields, the SP network enables readers to find matches to their interests at the finest granularity – the individuals whose work matches their own interests. For comparison, consider the large volume of e-mail I receive from journals sending me lists of their tables of contents. These e-mails are simply spam; essentially all the paper titles are of zero interest to me, so now I do not even bother to look at them. The subscription model only makes sense if it is *specific* to the subscriber's interests (otherwise he is better off just running a literature search). And in this day and age of highly specialized research, that means identifying *individual* authorities whose work matches your own.
- *Real metrics.* A key function of the SP network is to record all information about how each paper spreads through the community and to measure interest and opinions throughout this process. This will give readers detailed metrics about both reviewers (e.g., assessing their ability to predict what others will find interesting and important, ahead of the curve) and about papers (e.g., assessing not only their readership and impact but also how their level of interest spreads over different communities, and the community consensus on them, i.e., incorporated into the literature (via ongoing citations) or forgotten).
- *Higher quality.* Note first that the SPRs are simply the same referees that journals rely on, so the baseline reliability of their judgments is the same in either context. But the SP network aims for a *higher* level of quality and relevance – it only reports papers that are specially selected by referees as being of high interest to a particular subfield. "Ordinary research" (i.e., work that follows the pattern of work in its field) is typically judged by a standard "checklist" of technical expectations within its field. Unfortunately, a substantial fraction of such papers are technically competent but do not provide important new insights. The sad fact is that the average paper is only cited 1–3 times (over 2 years, even including *self-citations*), and indeed this distribution is highly skewed, in which the vast majority of papers have zero or very few citations, and only a small fraction of papers have substantial numbers of citations (Adler et al., 2008). For a large fraction of papers, the verdict of history is that almost *nobody* would be affected if these papers had *not* been published; even *their own authors* rarely get around to citing them!

Since the SP network is driven solely by *individual interest* (i.e., an SPR getting excited enough about a manuscript to recommend it to his subscribers), it is axiomatic that it will filter out papers that are not of interest to anyone. Since such papers unfortunately constitute a substantial fraction of publications, this is highly valuable service. A more charitable (but scarier) interpretation is that actually some fraction of these papers *would* be of interest to someone, but due to the inefficiencies of the journal system as a method for matching papers to readers, simply never find their proper audience. The SP network could “rescue” such papers, because it provides a fine-grained mechanism for small, specialized interest groups to find each other and share their discoveries.

- **Better information.** In a traditional journal, a great deal of effort is expended to critically review each manuscript, but when the paper is published, all of that information is discarded; readers are not permitted to see it. By contrast, in the SP network the review process is open and visible to all readers; the concerns, critiques and key tests of the paper’s claims are all made available, giving readers a much more complete understanding of the questions involved. Indeed, one good use for the SP network would be for reviewers and/or authors to make public the reviews and responses for papers published in traditional journals.
- **Speed.** When a new area of research emerges, it takes time for new journals to cover the new area. By contrast, the SP network can cover a new field from the very day that reviewers in its network start declaring that field in their list of interests. Similarly, the actual decision of a reviewer to recommend a paper can be fast: if they feel confident of their opinion, they can do so immediately without anyone else’s approval.
- **Long-term evaluation.** In a traditional journal, the critical review process ends weeks to months before the paper is published. In the SP network, that process continues as long as someone has something to say (e.g., new questions, new data) about that paper. The SP network provides a standard platform for everyone to enter their reviews, issues, and data, on papers at every stage of the life cycle.
- **Liberate referees to focus on their interests.** The SP network would urge referees to refuse to review anything that does not grab their interest, for the simple reason that it is both inefficient and counter-productive to do so. If a paper is not of interest to the referee, it is probably also not of interest to his subscribers (who chose his list because his interests match theirs). Note that the SP network expects authors to “submit” their manuscript simultaneously to multiple reviewers seeking an “audience” that is interested in their paper. If the authors literally cannot find anyone who *wants* to read the paper, it should not be recommended by the SP network. Note that if referees simply follow their own interests, this principle is enforced automatically.
- **Referees earn reputation and influence through their reviews.** Manuscript reviews are a valuable contribution to the research community, and they should be treated and valued as such. By establishing a record of fair, insightful reviews, and recommending important new papers “ahead of the curve,” a referee will attract a large audience of subscribers. This in and of itself should be treated as an important metric for professional evaluation. Moreover, the power to communicate directly with a substantial audience in your field itself constitutes *influence*, and is an important professional advantage. For example, a referee by default will have the right to communicate his own papers to his subscribers; thus, through his earned reputation and influence, a referee builds an audience for his own work.
- **Eliminate the politics of refereeing.** Note that a traditional journal does not provide referees these benefits because their role is fraught with the political consequences of acting as the journal’s “agent,” i.e., the power to confer or deny the right of publication, so crucial for academics. These political costs are reckoned so serious that journals shroud their referees in secrecy to protect them from retribution. Unfortunately, this political role incurs many other serious costs (see for example the problem of “prestige battles” analyzed in section 3.3).

These problems largely vanish in an SP network, for the simple reason that each referee represents no one but himself, and is *not* given arbitrary power to block publication of anyone’s work. In many traditional journals, if one reviewer says “I do not like this paper,” it will be rejected and the authors must start over again from scratch (since they are permitted to submit to only one journal at a time, and the paper must typically be re-written, or at a minimum re-formatted, for submission to another journal). By contrast, in an SP network authors submit their paper simultaneously to multiple referees; if one referee declines to *recommend* it, that has no effect on the other referees. The referee has not “taken anything away” from the authors, and has no power to block the paper from being selected by other referees.

Moreover, the very nature of the “Selected-Papers” idea is positive, that is, it highlights papers of especial interest for a given community. Being “selected” is a privilege and not a right, and is *intended* to reflect each referee’s idiosyncratic interests. Declining to select a paper is not necessarily a criticism; it might simply mean that the paper is not well-matched to that reviewer’s personal interests. Since most people in a field will also themselves be reviewers, they will understand that objecting to someone else’s personal selections is morally incompatible with preserving their own freedom to make personal choices.

### 3.2. BENEFITS FOR REFEREES

*Referees get all the disadvantages and none of the benefits of their own work in the current system.* Journals ask referees to do all the actual work of evaluating manuscripts (for free), but keep all the benefit for themselves. That is, if the referee does a good job of evaluating a manuscript, it is the *journal’s* reputation that benefits. This is sometimes justified by arguing that every scientist has an inherent obligation to review others’ work, and that failure to do so (for example, for a manuscript that has no interest to the referee) injures the cooperative enterprise of science. This is puzzling. Why should a referee *ever* review a paper except because of its *direct* relevance to his own work? If the authors (and the journal) cannot find *anyone* who actually *wants* to read the paper, what is the purpose of publishing it?

Reviewing manuscripts is an important contribution and should be credited as such. The SP network would rectify this in two ways:

Note that standard etiquette will be that authors may submit a paper to as many referees as they like, but at the same time referees are not obligated to respond.

Of course, in certain cases a referee may feel that important concerns have been ignored, and will raise them by publishing a *negative review* on his SP list. I believe that referees will feel free to express such concerns in this open setting, for the same reasons that scientists often speak out with such concerns at public talks (e.g., at conferences). That is, they are simply expressing their personal opinions in an open, public forum where everyone can judge the arguments on their merits. They are only claiming *equal* rights as the authors (i.e., the right to argue for their position in a public forum). What creates conflict in peer review by traditional journals is the fact that the journal gives the referee *arbitrary power* over the authors' work – specifically, to suppress the authors' right to present their work in a public forum. This power is made absolute in the sense that it is exercised in secret; the merit of the referees' arguments are not subject to public scrutiny; and referees have no accountability for whether their assertions prove valid or not. All of these serious problems are eliminated by the SP network, and replaced by the benefits of openness, transparency, and accountability.

- *Eliminate the costs of delegated review:* currently, researchers are called upon to waste significant amounts of time reviewing papers that are not of direct interest to their own work. Typically, this time constitutes a cost with no associated gain. By contrast, time spent reviewing a paper that is of vital interest for the referee's research gives him immediate benefit, i.e., early access to an important advance for his own work.

### 3.3. BENEFITS FOR AUTHORS

I now consider the benefits of the SP network review and publication system in terms of readership and cost. These benefits arise from addressing fundamental inefficiencies: first, how poorly traditional journals fit the highly specialized character of research and the emergence of new fields; and second, how journals have implemented peer review. Criticisms of this peer review system are legion, and most tellingly, come from inside the system, from Editors and reviewers (see for example Smith, 2009). While assessment of its performance is generally blocked by secrecy, the studies that have been done are alarming. For example, re-submission of 12 previously published articles was not detected by reviewers in 9 out of 12 cases (showing that reviewers were not familiar with the relevant literature), and 8 of the 9 papers were rejected (showing a nearly total lack of concordance with the previous set of reviewers who published these articles; Peters and Ceci, 1982). While we each can hope that reviewers in our own field would do better, there is evidently a systemic problem. That is, the system itself promulgates a high level of errors. I now argue that the SP network can help systematically address some of these errors.

#### 3.3.1. Readership

The SP network can help alleviate bottlenecks that impede publishing innovative work in the current system, e.g., because its specialized audience does not “have its own journal,” or because it is “too innovative” or “too interdisciplinary” to fare well in EPR. Let

us consider the case of a paper that introduces a novel combination of two previously separate expertises. In a traditional journal, the paper would be “delegated” to two or three referees who have not been chosen on the basis of a personal interest in its topic. So the probability that they can understand its significance for its target audience is low. For each of these referees, approximately half of the paper goes outside their expertise, and may well not follow the assumptions of their own field. Since they lack the technical knowledge to even evaluate its validity, the probability that they will feel *confident* in its validity is low. Even if the authors get lucky, and one referee ranks it as both interesting and valid, traditional “false-positive” screening requires that *all three* reviewers recommend it. Multiplying three poor probabilities yields a low probability of success. In practice, this conservative criterion leads to conservative results: it selects what “everybody agrees is acceptable.” It rewards staying in the average referee's comfort zone, and penalizes innovation.

By contrast, the SP network explicitly searches for interest in the paper, over a far larger number of possible referees (say 10–50), using fast, automatic click-through metrics. Obviously, if no one is interested, the process just ends. But if the paper is truly innovative, the savviest people in the field will likely be intrigued. Next, the interested reviewers question the authors about points of confusion, prior to stating any judgment about its validity. Instead of requiring *all* referees to recommend the paper for publication, the SP network will “publish” a paper if just one referee chooses to recommend it. (Of course in that case it will start out with a smaller audience, but can grow over time if any of those subscribers in turn recommend it). A truly innovative, sound paper is likely to get multiple recommendations in this system (out of the 10 or more SPRs to whom it was initially shown). By contrast with traditional publishing, it is optimized for a low false-negative error rate, because it selects what at least one expert says is extraordinary (and allocates a larger audience in proportion to the number of experts who say so). Any reduction the SP network makes in this false-negative rate will produce a dramatic increase in coverage for these papers.

#### 3.3.2. Cost

The SP network reduces the costs of publishing to the community (in terms of human time and effort) in several ways:

- *it eliminates the costs of “restart at zero” and the “non-compete clause”:* markets work efficiently only to the extent they actually function as free markets, i.e., via competition. It is worth noting that while papers compete to get into each journal, journals *do not compete* with each other for each paper. Journals enforce this directly via a “non-compete clause” that simply makes it illegal for authors to submit to more than one journal, and indirectly via incompatible submission systems and incompatible format requirements (even though there is little point applying such requirements until after the journal has decided to accept the paper). In practice an author must “start over from scratch” by re-writing and re-submitting his paper to another journal. Note that this multiplies the publication cost ratio for a paper by the number of times it must be submitted. It is not uncommon for this to double or triple the publication cost ratio.

From the viewpoint of the SP network, these “restart at zero” strictures are wasteful and illogical. On the one hand, it means that each editor gets only a small slice of the total review information (since the different reviews are kept separate, rather than pooled). On the other hand, it wastes an immense amount of time re-reviewing the same paper over and over. Finally, the SP network pools the parallel review efforts of all interested SPRs in a single unified process. Each SPR sees the complete picture of information from all SPRs, but makes his own independent decision.

Let us consider the publication cost ratios for different cases. For a paper that is not of strong interest to any audience, traditional journal review typically involves months of “restart at zero” reviews. By contrast, the SP network will simply return the negative result in a few days (“no interested audiences found”). Thus, the SP network reduces the publication cost ratio in this case by at least a factor of ten. For papers that require extensive audience search (either because they are in a specialized sub-field, or because they contain “too much innovation” or “too many kinds of expertise”), they again are likely to fall into the trap of “restart at zero” re-review, consuming months, and possibly yielding no publication. In the SP network, the authors should be able to find their audience (possibly small) within days, and then go through a single review process leading to publication by one or more SPRs. Because “restart at zero” is avoided, the publication cost ratio should be two to three times less. Finally, for papers with an obvious (easy to find) audience, the SP network still offers some advantages, basically because it guarantees a *single round of review*. By contrast, traditional peer review requires unanimity. This unavoidably causes a significant false-negative rate. Under the law of “restart at zero,” this means a certain fraction of good papers waste time on multiple rounds of review. For this category overall, I expect the publication cost ratio of traditional publishing to be 1–2 times that of the SP network.

- *it eliminates the costs of “gambling for readership”*: when researchers discover a major innovation or connection between fields, they become ambitious. They want their discovery published to the largest possible audience. Under the non-compete clause, this means they must take a gamble, by submitting to a journal with a large readership and correspondingly high rejection ratio. Often they start at the top (e.g., a *Nature* or *Nature Genetics* level journal) and work their way down until the paper finally gets accepted. Summed over the entire research community, this law of “restart at zero” imposes a vast cost with no productive benefit, i.e., the paper gets published regardless. The SP network avoids this waste, by providing an efficient way for a paper’s readership to grow *naturally*, as an automatic consequence of its interest to readers. Neither authors nor referees have to “gamble” on predictions of how much readership the paper should be “allocated.” Instead, the paper is simply released into the network, where it will gradually spread, in direct proportion to how many readers it interests.
- *it eliminates the costs of “prestige battles”*: referees for traditional journals play two roles. They explicitly assess the technical validity of a paper, but they also (often implicitly) judge whether it

is “prestigious enough” for the journal. Often referees decide to reject a paper based on prestige, but rather than expressing this subjective judgment (“I want to prevent this paper from being published here”), they justify their position via apparently objective criticisms of technical validity details. The authors doggedly answer these criticisms (often by generating new data). If the response is compelling, referees will commonly re-justify their position simply by finding new technical criticisms. Unfortunately, this process often doubles or triples the review process, and is unproductive, first because the referee’s decision is already set, and second because the “technical criticisms” are just red herrings; answering them does not address the referee’s real concern. Even if the paper is somehow accepted (e.g., the editor intervenes), this will double or triple the publication cost ratio.

By contrast, in the SP network this issue does not even arise. This problem is a pathology of non-expert PUSH – i.e., asking referees to review a paper they are not personally interested in. In the SP network, there is no “prestige factor” for referees to consider at all (first because the SP network simply *measures* impact long-term, and second because that metric has little dependency on what any individual reviewer decides). Indeed, the only decision a referee needs to make initially is whether they are personally interested in the paper or not. And that decision is measured instantly (via click-through metrics), rather than dragged out through weeks or months of arguments with the authors.

## 4. PRECEDENTS

This proposal is hardly original—it merely synthesizes what many scientists have argued for in a wide variety of forums (Hitchcock et al., 2002; Neylon, 2005; Nielsen, 2008; Kriegeskorte, 2009; Smith, 2009; Baez et al., 2010; The Peer Evaluation Team, 2010; Birukou et al., 2011; von Muhlen, 2011). There is powerful precedent for both a public publishing service, and for a recommendations-based distribution system. For example, arXiv is the preeminent preprint server for math, physics, and computer science (Cornell University Library, 1996). A huge ecology of researchers are using it as a *de facto* publishing system; it provides the real *substance* of publishing (lots of papers get posted there, and lots of people read them) without the official imprimatur of a journal.

As usual with such things, the main barrier to realizing the benefits of a new system is simply the entrenchment of the old system. In my view, the advantage of this proposal is that it provides a seamless bridge between the old and new, by working equally well with either. In the context of the old system, it is a social network in which everyone’s recommendations of published papers can flow efficiently. But the very act of using such a network creates a new context, in which every user becomes in a sense as important a “publication channel” as an established journal (at least for his subscribers).

### 4.1. EXAMPLES THAT AN SP NETWORK COULD BUILD UPON

In my view, most of the key ingredients are in place; what is needed is to integrate them together as an SP network. Here are some examples, by no means comprehensive:

- *Online bibliography managers* such as Academia.edu (The Academia.edu Team)<sup>1</sup>, CiteULike (Cameron et al., 2004), Connotea (Nature Publishing Group)<sup>2</sup>, Mendeley (The Mendeley Team)<sup>3</sup>, and ResearchGate (The ResearchGate Team)<sup>4</sup>. These provide public sites where researchers can save citations, rate papers, and share their ratings. CiteULike also attempts to recommend articles to a user based on his citation list. In principle, users' lists of favorite papers could be used as a source of recommendations for the SP network.
- *open peer review platforms*: PeerEvaluation.org has launched an open access manuscript sharing and open peer review site (The Peer Evaluation Team, 2010). Peer review is open (non-anonymous), and it also seeks to provide "qualitative metrics" of impact. It could be viewed as a hybrid of arXiv (i.e., an author self-publishes by simply uploading his paper to the site) plus open, community peer review.
- *improved metrics*: The LiquidPub Project analyzed a wide variety of metrics for assessing impact and peer review quality; for a review see Birukou et al. (2011).
- *journals that support aspects of open peer review*: PLoS ONE (Public Library of Science, 2006) represents an interesting precedent for the SP network. In terms of its "back-end," PLoS ONE resembles some aspects of the SP network. For example, its massive list of "Academic Editors" who each have authority to accept any submitted paper is somewhat similar to the "liberal" definition of SPRs that allows any SPR to recommend a paper to his subscribers. However, on its "front-end" PLoS ONE operates like a traditional journal: reviews are secret; no effort is made to search for a paper's audience(s); and above all there is no network structure for papers to spread naturally through a community.

*Biology Direct* (Koonin et al., 2006) is another interesting precedent. It employs a conventional (relatively small) editorial board list. However, like the SP network, it asks authors to contact possible reviewers from this list *directly*, and reviewers are encouraged to decline a request if the paper does not interest them. Moreover, reviews are made public when a paper is published. Again, however, *Biology Direct's* front-end is that of a conventional journal, with no network structure.

## 4.2. LESSONS FROM THESE PRECEDENTS

Given that these sites already provide important pieces of this proposal, it is interesting to ask why they have not already succeeded in creating an SP network. I see two basic reasons:

- Several pieces must be put together before you have a *network* that can truly act as content distribution system. For example, people do not normally think of bibliography management (e.g., CiteULike) as a distribution system, and there are good reasons for this. Bibliography managers do not solve the fundamental problems of publication, namely *audience search*

(finding a channel that will reach the audience of people that would read the paper), *validation* (identifying all issues which could undercut the paper's claims, and figuring out how to address them), and *distribution* (actually reaching the audience). There are certainly aspects of CiteULike, Mendeley, etc., that could be applied to solve the distribution problem (e.g., paper recommendations), but this will not happen until all of the components are present and working together.

- These sites are "yet another thing" a busy scientist would have to do (and therefore is unlikely to do), rather than something that is integrated into what he *already* does. For example, I think that a scientist is far more likely to view (and make) recommendations linked on PubMed search result pages, than if we ask him to log in to a new website such as CiteULike. The problem is the poor balance of incentives vs. costs for asking the scientist to use a new website: on the one hand, any recommendations he makes are unlikely to be seen by many people (because a new site has few users); on the other hand, he has to go out of his way to remember to use the site.

To create a positive balance of incentives vs. costs, an SP network must (a) make reviewing truly important (i.e., it must gate whether the paper gets published, just like peer review at a journal); (b) reward reviewers by prominently displaying their recommendations directly on PubMed search results and the journal website, etc. (so that recommendations you write will be immediately seen by many readers, even if you do not yet have any subscribers); (c) make it easy for all scientists to start participating, directly from sites they already use (such as PubMed). For example, a page showing a paper at PubMed or the journal website should have a "Like!" link that enables the reader to enter a recommendation directly; (d) help authors search for the specific audience(s) for their paper through automated click-through metrics. This harnesses a real motive force – the quest for your personal scientific interests, both as an author and a reader – in service of getting people to participate in the new peer review system.

These precedents also suggest that an SP network should be open to a wide variety of communication methods – by providing a common interface that many different sites could plug in to – rather than trying to create a single site or mechanism that everyone must use. Ideally, all of these different sites (e.g., CiteULike, PeerEvaluation.org) should be able to both view and enter information into the SP network. In this way, the SP network serves to tie together many different efforts. For example, it might be possible to create mechanisms for the SP network to draw from the large number of researchers who are using blogs to discuss and review their latest finds in the literature, some of them are extremely influential (e.g., Tao, 2007), and John Baez/n-category cafe (Baez, 1993–2010; Baez et al., 2007), to cite two examples). As one simple example of allowing many input methods, the SP network should make it easy for a blog user to indicate which of his blog posts are reviews, and what papers they recommend, automatically delivering these recommendations to his subscribers.

In my view, it is very important that the SP network be developed as an open-source, community project rather than as a commercial venture, because its data are freely provided by the research community, and should be freely used for its benefit.

<sup>1</sup><http://academia.edu/>

<sup>2</sup><http://connotea.org/>

<sup>3</sup><http://www.mendeley.com/>

<sup>4</sup><http://researchgate.net/>



To the extent that they become valuable, commercial sites tend to become “walled gardens” in which the community is encouraged to donate content for free, which then becomes the property of the company. That is, it both *controls* how that content can be used, and uses that content for *its own benefit* rather than that of the community. The SP network would provide enormous benefits to the community, but from the viewpoint of a publishing company (e.g., NPG) it might simply look like a threat to their business. The SP network should be developed as a walled garden, because its data belong to the community and must be used for the community’s benefit. It must be developed “of the people, by the people, for the people,” or it will never come to *be* in the first place.

## 5. A THREE PHASE PLAN FOR BUILDING A SELECTED-PAPERS NETWORK

To provide concrete details about how this concept could work, I outline how it could be implemented in three straightforward, practical phases:

- Phase I: *the basic SP network*. Building a place where reviewers can enter paper selections and post-reviews, readers can search and subscribe to reviewers’ selections, and papers’ diffusion through research communities is automatically measured.
- Phase II: *A better platform for scientific publishing*. This phase will focus on providing a comprehensive platform for open peer review, as an alternative to journals’ in-house peer review. Authors would be invited to submit directly to the SP network peer review platform, and then after its review process was complete, to invite a journal editor to decide whether to accept the paper on the basis of the SPR’s reviews. To make this an attractive publishing option, it will give authors powerful tools for quickly locating the audience(s) for a paper, and it will give reviewers powerful tools for pooling expertise to assess its validity, in collaboration with the authors. All of this is driven by the SP network’s ability to target specialized audiences far more accurately, flexibly, and quickly than traditional journals. One way of saying this is that the SP network automatically creates a new “virtual journal” (list of subscribers) optimized for each individual paper, and that this is done in the most direct, natural way possible (i.e., by each reviewer deciding whether or not to recommend the paper to his subscribers). Note that this strategy aims not at supplanting traditional journals but complementing them. This alternative path will be especially valuable for specialized subfields that are not well-served by existing journals, for newly emerging fields, and for interdisciplinary research (which tends to “fall between the cracks” of traditional journal categories).
- Phase III: *discovering and measuring the detailed structure of scientific networks*. I propose that the SP network should record not only of the evolution of the subscription network (revealing sub-communities of people who share a common interest as shown by cliques who subscribe to each other), but also the exact path of how each paper spreads through the network. Together with a wide range of automatic measurements of each reader’s interest in a paper, these data constitute a golden opportunity for rigorous research on knowledge networks and social networks (e.g., statistical methods for discovering the creation

of new subfields directly from the network structure). Properly developed, this dataset would enable new scientometrics research and will produce a wide variety of new algorithms (e.g., Netflix-style prediction of a paper’s level of interest for any given reader) and new metrics (e.g., how big is a reviewer or author’s influence within his field? How accurately does he predict what papers will be of interest to his field, or their validity? How far “ahead of the curve” is a given reviewer or author?). Note that the SP network needs only to capture the data that *enables* such research; it is the research community that will actually do this research. But the SP network then benefits, because it can put all these algorithms and metrics to work for its readers, reviewers and authors. For example, it will be able to create publishing “channels” for new subfields as soon as new cliques are detected within the SP network structure.

### 5.1. PHASE I: BUILDING A SELECTED-PAPERS NETWORK

Technically, the initial deployment requires only a few basic elements:

- *a mechanism for adding reviewers* (“Selected-Paper Reviewer” or SPR): the SP network restricts reviewers in a field simply to those who have published peer reviewed papers in that field (typically as corresponding author). Initially it will focus on building (by invitation) a reasonably comprehensive group of reviewers within certain fields. In general, any published author from any field can add themselves as a reviewer by linking their e-mail address to one of their published papers (which usually include the corresponding author’s e-mail address). Note that the barrier to entry need not be very high, since the only privilege this confers is the right to present one’s personal recommendations in a public forum (no different than starting a personal blog, which anyone can do). Note also that the initial “field definitions” can be very broad (e.g., “Computational Biology”), since the purpose of the SP network is to enable sub-field definitions to emerge naturally from the structure of the network itself.
- *a mechanism for publishing reviews*: Peer reviews represent an important contribution and should be credited as such. Concretely, substantive reviews should be *published*, so that researchers can read them when considering the associated paper; and they should be *citable* like any other publication. Accordingly, the SP network will create an online journal *Critical Reviews* that will publish submitted reviews. The original paper’s authors will be invited to check that a submitted review follows basic guidelines (i.e., is substantive, on-topic, and contains no inappropriate language or material), and to post a response if desired. Note that this also triggers inviting the paper’s corresponding author to become an SP reviewer (by virtue of having published in this field). Reviews may be submitted as *Recommendations* (i.e., the reviewer is selecting the paper for forwarding to his subscribers), *Comments* (neutral: the review is attached to the paper but not forwarded to subscribers), or *Critiques* (negative: a warning about serious concerns. The reviewer can opt to forward this to his subscribers). *Recommendations* should be written in “News and Views” style, as that is their function (to alert readers to a

potentially important new finding or approach). *Comments* and *Critiques* can be submitted in standard “Referee Report” style. Additional categories could be added at will: e.g., *Mini Reviews*, which cover multiple papers relevant to a specific topic (for an excellent example, see the blog *This Week’s Finds in Mathematical Physics*; Baez, 1993–2010); *Classic Papers*, which identify must-read papers for understanding a specific field; etc.

Note also that the SP network can give reviewers multiple options for how to submit reviews: via the Science Select website (the default); via Google Docs; via their personal blog; etc. For example, a reviewer who has already written “News and Views” or mini reviews on his personal blog, could simply give the SP network the RSS URL for his blog. He would then use the SP network’s tools (on its website) to select the specific post(s) he wants to publish to his subscribers, and to resolve any ambiguities (e.g., about the exact paper(s) that his review concerns).

- *a subscription system*: the SP network would define an open standard by which any site that displays paper titles, abstracts, or full-text could link to the ranked set of recommendations for those papers, or let its users easily make paper recommendations. For example, the PubMed search engine could display a “Recommended” link next to any recommended paper title, or (when displaying an abstract) a ranked list of people who recommended the paper. In each case these would be linked to that person’s review of the paper, their other paper recommendations, and the option to subscribe to their future recommendations. Similarly, it would display a “Like!” icon that would let the user recommend the paper. This would give people a natural way to start participating immediately in the SP network by viewing and making recommendations anywhere that they view papers – whether it be on PubMed, a journal’s website, etc.

Subscribers could opt to receive recommendations either as individual e-mails; weekly/monthly e-mail summaries; an RSS feed plugged into their favorite browser; a feed for their Google Reader; or other preferred news service, etc. Invitations will emphasize the unique value of the SP network, namely that it provides the subscriber *reviews* of important new papers specifically in his area (whereas traditionally review comments are hidden from readers).

- *an automatic history-tracking system*: each paper link sent to an individual subscriber will be a unique URL, so that when s/he accesses that URL, the system will record that s/he viewed the paper, as well as the precise *path* of recommenders via which the paper reached this reader. In other words, whereas the stable internal ID for a paper will consist of its DOI (or arXiv or other database ID), the SP network will send this to a subscriber as a URL like <http://doc.scienceselect.net/Tase3DE6w21>. . . that is a unique hash code indicating a specific *paper* for a specific *subscriber*, from a specific *recommender*. Clicking the title of the paper will access this URL, enabling the system to record that this user actually viewed this paper (the system will forward the user to the journal website for viewing the paper in the usual way). If this subscriber then recommends the paper to his own subscribers, the system sends out a new set of unique links and the process begins again. This enables the system to track the

exact path by which the paper reached each reader, while at the same time working with whatever sources the user must access to actually read any given paper. Of course, the SP network will take every possible measure to prevent exposure or misuse of these data. These metrics should include appropriate controls for excluding trivial effects such as an attention-grabbing title. Since the SP network directly measures the probability that someone will recommend the paper after reading it, it should be able to control for such trivial effects.

- *an automatic interest-measuring system*: click-through rates are a standard measure of audience response in online advertising. The SP network will automatically measure audience interest via click-through rates, in the following simple ways:
  - The system will show (send) a user one or more paper titles. The system then measures whether the user clicks to view the abstract or review.
  - The system displays the abstract or review, with links to click for more information, e.g., from the review, to view the paper abstract or full-paper. Each of these click-through layers (title, review, abstract, full-paper) provides a stronger measure of interest.
  - The system provides many options for the user to express further interest, e.g., by forwarding the paper to someone else; “stashing” it in their personal cubbyhole for later viewing; rating it; reviewing or recommending it on their SP list, etc.
- *a paper submission mechanism*: while reviewers are encouraged to post-reviews on their own initiative, the SP network will also give authors a way to invite reviews from a targeted set of reviewers. Authors may do this either for a published paper (to increase its audience by getting “selected” by one or more SPRs, and spreading through the SP subscriber network), or for a preprint. Either way, authors must supply a preprint that will be archived on the SP network (unless they have already done so on standard repository such as arXiv). This both ensures that all reviewers can freely access it, and guarantees Open Access to the paper (the so-called “Green Road” to open access). (Note that over 90% of journals explicitly permit authors to self-archive their paper in this way; Harnad et al., 2004). Authors use the standard SP subscriber tools to search for relevant reviewers, and choose up to 10 reviewers to send the paper link to. Automatic click-through measurements (see section below) will immediately assess whether each reviewer is interested in the paper; actually proceeding to read the paper (“whoa! I gotta read this!”) triggers an invitation to review the paper. These automatic interest metrics should be complete within a few days. For reviewers who exhibit interest in a paper, the authors follow up with them directly. As always, each reviewer decides at their sole discretion whether or not to recommend the paper to their subscribers. As in traditional review, a reviewer could demand further experiments, analysis, or revisions as a condition for recommending the paper. While each reviewer makes an independent decision, all reviewers considering a paper would see all communications with the authors, and could chime in with their opinions during any part of that discussion.

It is interesting to contrast SP reviewer invitations vs. the constant stream of review requests that we all receive from journals. While

SP reviewers could in principle receive a larger number of “paper title invitations,” this imposes no burden of demands on them; i.e., *no one is asking them to review anything unless it is of burning interest to them*. There is no nagging demand for a response; indeed, reviewers will be expressly instructed to ignore anything that does not grab their interest!

## 5.2. PHASE II: THE SP NETWORK AS A PEER REVIEW PLATFORM

The capabilities developed in phase I provide a strong foundation for giving authors the choice of submitting their work directly to the SP network as the peer review mechanism (which could result in publication in a traditional journal). To do this, the SP network will make these capabilities available as a powerful suite of tools 1. for authors to search for the audience(s) that are interested in their work; 2. for authors and referees to combine their different expertises (in synthesis rather than opposition) to identify and address key issues for the paper’s impact and validity; 3. for long-term evaluation after a paper’s publication, to enable the community to raise new issues, data, or resolutions. This will be particularly useful for newly emerging fields (which lack journals) or subfields that are not well-served by existing journals.

However, it must be emphasized that this is not an attempt to compete with or replace traditional journals. Instead, the SP network complements the strengths of traditional journals, and its suite of tools could be useful for journals as well. Concretely, the SP network will develop its tools as an open-source project, and will make its software and services freely available to journals as well as to the community at large. For example, journals could use the SP network’s services as their submission and review mechanism, to gain the many advantages it offers over the very limited tools of traditional review (which consist of little more than an ACCEPT/REJECT checkbox for the Editor, and a text box for feedback to the authors).

### 5.2.1. The SP network publication process

The SP network will provide tools for “market research” (i.e., finding the audience(s) that are interested in a given paper) and for synthesis (integrating multiple expertises to maximize the paper’s value for its audience(s)), culminating in publication of a final version of the paper (by being selected by one or more SPRs). I will divide this into three “release stages”: alpha (market research); beta (synthesis); post-publication (long-term evaluation). These are analogous to the alpha-testing, beta-testing, and post-release support stages that are universal in the software industry. The alpha release cycle identifies a specific audience that is excited enough about the paper to work on reviewing it. The beta release cycle draws out questions and discussion from all the relevant expertise needed to evaluate the paper and optimize it for its target audience(s). The reviewers and authors work together to raise issues and resolve them. Individual reviewers can demand new data or changes as pre-conditions for recommending the paper on their SP list. On the one hand, the authors decide when the paper is “done” (i.e., to declare it as the final, public version of the paper). On the other hand, each reviewer decides whether or not to “select” the paper for their SP list. On this basis, authors and reviewers negotiate throughout the beta period what will go into the final release. As long as one SPR elects to recommend the paper to his

subscribers, the authors have the option of publishing the paper officially in the SP network’s journal (e.g., *Selected-Papers in Biology*). Regardless of how the paper is published, the same tools for synthesis (mainly an *issue tracking system*) will enable the entire research community to continue to raise and resolve issues, and to review the published paper (i.e., additional SPRs may choose to “select” the paper).

### 5.2.2. Alpha release tools

For alpha, the tools already provided by Phase I are sufficient: e.g., the *paper submission mechanism*; methods for measuring *reader interest*; and *audience search* methods. Here I will simply contrast it with traditional peer review.

- *assess impact, not validity*: I wish to emphasize that alpha focuses entirely on measuring the paper’s impact (interest level) over its possible audiences. It does *not* attempt to evaluate the paper’s validity (which by contrast tends to dominate the bulk of referee feedback in traditional peer review). There are three reasons. First, impact is the key criterion for the SP network: if no SPR is excited about the paper, there is no point wasting time assessing its validity. Second, for papers that combine multiple expertises, its *impact* might lie within one field, yet it might use methods from another field. In that case, a referee who is expert in evaluating the *validity* of the methodology would not be able to assess the paper’s impact (which lies outside his field). Therefore in IDPR impact must often be evaluated separately. Third, the SP network is very concerned about *failing* to detect papers with truly novel approaches. Such papers are both less common, and harder for the average referee to understand in their entirety. This makes it more likely that referees will feel doubt about a novel approach’s validity. To avoid this serious risk of false-negatives, the SP first searches for SPRs who are excited about a paper’s potential impact, completely separate from assessing its validity.
- *impact-driven review, not non-expert PUSH*: traditional journals do essentially nothing to help authors find their real target audience, for the simple reason that journals have no tools to do this. Exploring the space of possible audiences requires far more than a single, small sample (2–3 reviews). It requires efficiently measuring the interest level from a meaningful sample for each audience. The key is that the SP network will directly *measure* interest (see the metrics described in Phase I and Phase III) over multiple audiences. By contrast, non-expert PUSH tends to produce high false-negative rates, because people are not good at predicting the interest level of papers that they themselves are not interested in. Being unaware of a paper’s interest for a problem outside your knowledge, and being unaware that another group of people is interested in that problem, tend to go together.
- *speed*: because alpha requires no validity review, it can be fast and automatic. The SP network’s click-through metrics can be measured for 10–100 people over the course of just a few days; advertisers (e.g., Google) measure such rates over vastly larger audiences every day.
- *journal recommendation system*: whenever a researcher expands the scope of his work into a new area, he initially may be unsure

where to publish. The SP network can automatically suggest appropriate journals, by using its interest measurement data. Simplistically, it can simply relate the set of SPRs who expressed strong interest in the paper to the set of journals which published papers recommended by those same SPRs.

### 5.2.3. Beta release tools

Beta consists of several steps:

- **Q & A:** This means that reviewers with different relevant expertise raise questions about the paper, and work with the authors to resolve them, using an online *issue tracker* that makes it easy to see what issues have already been raised, their status, and detailed discussion. Such systems provide great flexibility for synthesizing a consensus that draws on multiple expertises. For example, one referee may resolve another referee's issue. (A methodology reviewer might raise the concern that the authors did not follow one of the standard assumptions of his field; a reviewer who works with the data source analyzed in the paper might respond that this assumption actually is not valid for these data). Powerful issue tracking systems are used universally in commercial and open-source software projects, because they absolutely need such synthesis (to find and fix all their bugs). Using a system that actually supports synthesis changes how people operate, because the system makes it obvious they are all working toward a shared goal. Note that such a system is like a structured wiki or "threaded" discussion in that it provides an open forum for anyone to discuss the issues raised by the paper.

The purpose of this phase is to allow referees to ask all the questions they have in a non-judgmental way—a conversation with the authors, and with the other referees—before they even enter the Validity Assessment phase. This should distinguish clearly several types of questions:

- False-positive: Might result/interpretation X be due to some other explanation, e.g., random chance; bias; etc.? Indicate a specific test for the hypothetical problem.
- False-negative: is it possible your analysis missed some additional results due to problem Y? Indicate a specific test for the hypothetical problem.
- Overlap: how does your work overlap previous study X (*citation*), and in what ways is it distinct?
- Clarification/elaboration: I did not understand X. Please explain.
- Addition: I suggest that idea X is relevant to your paper (*citation*). Could that be a useful addition?

Each referee can post as many questions as he wants, and also can "second" other referees' questions. Authors can immediately answer individual questions, by text or by adding new data/analyses. Referees can ask new questions about these responses and data. Such discussion is important for synthesis (combining the expertise of all the referees and the authors) and for definitive clarification. It should leave no important question unanswered.

- **validity assessment:** eventually, these discussions culminate in each reviewer deciding whether there are serious doubts about the validity of paper's data or conclusions. While each reviewer decides independently (in the sense that only he decides what to

recommend on his SP list), they will inevitably influence each other through their discussions.

- **improving the paper's value for its audience:** once the critical validity (false-positive) issues are resolved, referees, and authors should consider the remaining issues to improve the manuscript, by clarifying points that confused readers, and adding material to address their questions. To take an extreme example, if reviewers feel that the paper's value is obscured by poor English, they might demand that the authors hire a technical writer to polish or rewrite parts of it. Of course, paper versions will be explicitly tracked through the whole process using standard software (e.g., Git; Torvalds and Hamano, 2005).
- **public release version:** the authors decide when to end this process, and release a *final version* of the paper. Of course, this is closely tied to what the reviewers demand as conditions for recommending the paper.

### 5.2.4. Publication

Authors can use the SP network alpha and beta processes to demonstrate their paper's impact and validity, and then invite a journal editor to consider their paper on that basis. A journal editor can simply join the beta process for such a paper; like the other SPRs, he decides (based on the complete synthesis of issues and resolutions in the issue tracker) whether he wishes to "select" the paper. The only difference is that he is offering the authors publication in his journal, whereas the other referees are offering a recommendation on their SP lists. Of course, the paper will typically have to be re-formatted somewhat to follow the journal's style guidelines, but that is a minor issue; extra material that does not fit its size limits can be posted as an online Supplement.

Note that this process offers many advantages to the journal. It does *not* need to do any work for the actual review process (i.e., to find referees, nag them to turn in reviews on time, etc.). More importantly, it gets all of the SP network's impact measurements for the paper, allowing it to see exactly what the paper's level of interest is. Indeed, the journal can get a "free-ride" on the SP network's ability to market the paper, by simply choosing papers that multiple (or influential) SPRs have decided to recommend to their subscribers. If the journal decides to publish such a paper, all that traffic will come to *its* website (remember that the SP network just forward readers to wherever the paper is published). For a journal, the SP network is a gold mine of improved review process and improved marketing – all provided to the journal for free.

However, an even greater value of the SP network review system is for areas that are not well-served by journals. If an SPR selects a paper for recommendation to his subscribers, the authors can opt to officially publish the paper in the SP network's associated journal. Note that this serves mainly to get the paper indexed by search engines such as PubMed, and to give the paper an "official" publication status. After all, the real substance of publication is *readership*, and being recommended on the SP network already provides that directly.

## 5.3. PHASE III: ANALYSIS AND METRICS FOR SCIENTIFIC NETWORKS

Here I will only briefly list some basic metrics that the SP network will incorporate into the peer review process. Of course, data collected by the SP network would make possible a wide range of

scientometric analyses, far beyond the scope of this paper. There is a large literature exploring new metrics for impact; for a review see Birukou et al. (2011).

- *rigorously controlled and validated methodologies for automatic measurement of reader interest.* The basic SP network approach of dividing content into “access layers” (e.g., title; review; abstract; full-paper; etc.) and measuring click-through rates provides a foundation for automatic measurement of interest in a paper within specific audiences. However there are many questions about how best to “control” for various sources of

noise to produce a robust, uniformly normalized measure of interest. These are research questions and should be answered by experimentally testing different “control” methods and directly validating their results. As a trivial example, click-through rates can be artifactually depressed if an unusually large fraction of the target audience is “offline,” e.g., during holidays or a major conference in the discipline. Such artifacts can be eliminated by measuring interest *relative* to a consistent control, i.e., by including multiple titles in any test mailing, one of which would be a “control.” Different papers for a given audience would be measured relative to the same control during any given time

**Table 1 | Traditional peer review vs SP network.**

Traditional peer review	SP network
<b>Expert peer review (EPR):</b> assumes each referee is expert in all aspects of the paper.	<b>Interdisciplinary peer review (IDPR):</b> a paper may combine more than one expertise, and thus may need a mix of referees, each of whom may not be expert on all aspects of the paper.
<b>Non-expert PUSH:</b> 2 or 3 reviewers try to guess what everyone else in the world (with different interests and expertise) will be interested in. Takes weeks to months.	<b>Measured impact:</b> impact is directly measured over a broad audience of researchers from different possible target areas, via instant click-through metrics. Takes a few days.
<b>Journal=TV channel:</b> every paper in it reaches the same fixed mass audience. For any individual reader, only a small fraction of papers in the journal are of interest (i.e., he would choose to read them). This is because scientists specialize much more finely than journals can.	<b>Virtual journal</b> created for each paper via active audience search and each reviewer’s recommendation to his own subscribers. A reader subscribes only to reviewers who match his specific interests, so a high fraction of papers recommended by such a reviewer (based on his own interests) will interest that reader.
<b>Shoot first and ask questions later:</b> each reviewer is called on to make and state an initial ACCEPT/REJECT decision by himself, without any feedback about aspects of the paper that are outside his expertise.	<b>Synthesis (understanding) before judgment:</b> reviewers and authors collaborate to raise validation questions and discuss what assumptions and criteria are appropriate for assessing the paper, <i>before</i> trying to make any validity decision.
<b>One man, one nuke:</b> one reviewer can kill the paper.	<b>One man, one vote:</b> no one has power to block a paper; each reviewer separately decides whether to recommend the paper to his own subscribers.
<b>High false-negative rate:</b> innovative, boundary-crossing papers are <i>more</i> likely to be rejected due to IDPR errors.	<b>Low false-negative rate:</b> innovative, boundary-crossing papers are <i>more</i> likely to be recommended.
<b>High false-positive rate:</b> a large fraction of papers published by peer reviewed journals interest no one, as shown by lack of citations.	<b>Low false-positive rate:</b> a reviewer must find a paper of high interest, to recommend it to his subscribers.
<b>Restart at zero:</b> peer review is fragmented and wasteful because each journal ignores previous reviews and starts over at zero. The cost in time and effort for publishing a paper is multiplied by the number of journals the paper must be (re)submitted to.	<b>Unified review:</b> a single set of reviewers collaborates to review the paper and then make independent decisions about whether to recommend it to their subscribers. Journal editors decide based on those reviews (and the known initial audience size given by those recommendations) whether the paper is right for their journal’s audience. Each journal can see <i>all</i> the reviews; the paper never needs to be re-reviewed.
<b>The reviews are thrown away:</b> after the enormous effort involved in reviewing a paper, no one is permitted to see the reviews.	<b>Reviews are published:</b> the research community needs to see the important concerns and issues elucidated by the reviews. Referees should receive credit (if they want it) for this vital contribution.
<b>Referee protection program:</b> the review process is warped by the enormous political power each reviewer is burdened with (he must decide whether everyone else in the world should be allowed to see the paper).	<b>Speak for yourself:</b> no reviewer has the power to kill the paper, because everyone just decides for himself whether the paper is of interest to <i>him</i> (and makes no such judgment on anyone else’s behalf).
<b>Delegated review:</b> referees are repeatedly asked to waste time reviewing papers that are not of interest to their work (i.e., which they would not otherwise read).	<b>Interest-only review:</b> referees are instructed to refuse to review anything that they would not themselves <i>choose</i> to read (because of its compelling interest for their own work).



frame. Optimal signal-to-noise requires a control with a moderate interest level (neither too high nor too low), raising many interesting research questions about optimizing and automating these methods.

- *standardized measures of comparative interest for all papers.* Currently, the universal standard metric is simply the name of the journal in which the paper was published (i.e., “Nature” ≫ “Nucleic Acids Research” ≫ “unpublished preprint”). Many studies have shown that this “metric” is fatally flawed by huge variations in impact among papers published in the same journal (Adler et al., 2008). Another standard metric, citation impact, cannot be measured until two calendar years after publication, and thus is not useful during the period when readers need an interest metric (i.e., to guide their choice of what to read among recently published papers). Using its rigorous foundation of immediate interest metrics measured in real-time, the SP network can supply an important market need for a standardized measure of comparative interest that readers will intuitively understand. For example, since the SP network measures interest for all papers in the same, consistent way, it could report each paper’s interest level in terms of its “journal equivalent” by comparing the paper’s interest metric vs. the median interest metric for papers in a well-known journal. Note that by this measure some *Nature* papers might be reported as having an interest level equivalent only to an average *Nucleic Acids Research* paper, whereas some *Nucleic Acids Research* papers would be reported as having interest as high as an average *Nature* paper.
- *automatic “audience search” to identify the set of distinct audience(s) that would be interested in a specific new paper.* For a completely new paper, the system can predict its level of interest for different audiences, but its confidence intervals might be poor. By quickly measuring the actual interest in the most promising audiences (i.e., by showing the title to random samples of individuals from the target audience(s) and measuring click-through rates) it can both get more confident estimates for these audiences, and updated predictions for other audiences/individuals who are likely to be interested. Multiple cycles of this process can be run automatically over a timeframe of a few days, for example to give authors a validated list of target audience(s), among whom they could then ask reviewers to

consider their paper. Note that such methods would enable the SP network to auto-generate a “virtual journal” (unique list of subscribers) optimized for each specific paper. Whereas traditional journals function as purely “passive containers” with essentially static audiences, the SP network would gradually transform itself into an “active matrix” that uses rapid cycles of interest-prediction and online test-marketing to actively seek out the true audience(s) for each paper.

## CONCLUSION

One way of restating this proposal is that the challenges of scientific communication are too large for any one individual. In many fields, innovative papers tend to combine multiple expertises such that a referee will find some part of the paper goes outside his expertise. Yet standard peer review asks him to review it as if he were a universal expert able to decide both its impact and validity, by himself. The system places the whole burden of decision on a single individual (one referee can block the paper). It gives him no tools for sharing this burden by systematically collaborating with others with different expertise. It forces upon him an all-or-nothing distribution decision (ACCEPT or REJECT), because it lacks any way to break that decision down into finer granularity (e.g., different decisions for different sub-audiences).

In general, journal peer review suffers systemically from pathologies of excessive centralization, in other words, asking *one person* to make a decision for *everyone else*, when there is no sound basis for him to do so. The SP network solves these problems first by breaking them into many independent decisions distributed over many people, and second by integrating those people together with good tools for sharing expertise and collaborating in this assessment. **Table 1** summarizes how the SP network breaks down the tasks of peer review much more finely and effectively. This is a consummation devoutly to be wished.

## ACKNOWLEDGMENTS

The author wishes to thank Russ Altman, John Baez, Marc Harper, Nicholas Kriegeskorte, David Lipman, Cameron Neylon, Michael Nielsen, and the referees for their valuable critiques of this manuscript. This research was supported by the Office of Science (BER), U. S. Department of Energy, Cooperative Agreement No. DE-FC02-02ER63421.

## REFERENCES

- Acharya, A., and Verstak, A. (2005). Google Scholar. Available at: <http://scholar.google.com/>
- Adler, R., Ewing, J., and Taylor, P. (2008). *Citation Statistics. A Report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)*. IMU Joint Committee on Quantitative Assessment of Research. Available at: <http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf>
- Baez, J. (1993–2010). *This Week's Finds in Mathematical Physics*. Available at: <http://math.ucr.edu/home/baez/TWF.html>
- Baez, J., Corfield, D., Hoffnung, A., Huerta, J., Leinster, T., Shulman, M., Schreiber, U., and Willerton, S. (2007). *The n-Category Café*. Available at: <http://golem.ph.utexas.edu/category/>
- Baez, M., Birukou, A., Casati, F., and Marchese, M. (2010). Addressing information overload in the scientific community. *IEEE Internet Comput.* 14, 31–38.
- Birukou, A., Wakeling, J. R., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K., Osman, N., Ragone, A., Sierra, C., and Wasef, A. (2011). Alternatives to peer review: novel approaches for research evaluation. *Front. Comput. Neurosci.* 5:56. doi:10.3389/fncom.2011.00056
- Cameron, R., Hall, C., Emamy, K., and Caddy, J. (2004). *CiteULike.Org: A Free Service for Managing and Discovering Scholarly References*. Available at: <http://citeulike.org/>
- Cornell University Library. (1996). *The arXiv e-print Repository*. Available at: <http://arxiv.org/>
- Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H., and Hilf, E. R. (2004). The access/impact problem and the green and gold roads to open access. *Ser. Rev.* 30, 36–40.
- Hitchcock, S., Brody, T., Gutteridge, C., Carr, L., Hall, W., Harnad, S., Bergmark, D., and Lagoze, C. (2002). Open citation linking: the way forward. *D-Lib Mag.* 8.
- Koonin, E. V., Landweber, L. F., and Lipman, D. J. (2006). *Biology Direct, An Open Access, Peer-Reviewed Online Journal*. Available at: <http://www.biology-direct.com/>
- Kriegeskorte, N. (2009). *The Future of Scientific Publishing: Open Post-Publication Peer Review*. Available at: <http://futureofscipub.wordpress.com/>
- Lee, C. (2006). Peer review of interdisciplinary scientific papers: boundary-crossing research meets

- border patrol. *Nature*. Available at: <http://www.nature.com/nature/peer-review/debate/nature05034.html>
- National Library of Medicine. (1996). *The PubMed Database of Biomedical Literature*. Available at: <http://www.ncbi.nlm.nih.gov/pubmed>
- Neylon, C. (2005). *Science in the Open*. Available at: <http://cameronneylon.net/>
- Nielsen, M. (2008). *The Future of Science*. Available at: <http://michaelnielsen.org/blog/the-future-of-science-2/>
- Peters, D. P., and Ceci, S. J. (1982). Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behav. Brain Sci.* 5, 187–195.
- Public Library of Science. (2006). *PLoS ONE: Accelerating the Publication of Peer-Reviewed Science*. Available at: <http://www.plosone.org/>
- Smith, R. W. (2009). In search of an optimal peer review system. *J. Particip. Med.* 1, e13.
- Tao, T. (2007). *What's New*. Available at: <http://terrytao.wordpress.com/>
- The Peer Evaluation Team. (2010). *PeerEvaluation.Org: Empowering Scholars*. Available at: <http://peer-evaluation.org/>
- Torvalds, L., and Hamano, J. (2005). *Git: The Fast Version Control System*. Available at: <http://git-scm.com/>
- von Muhlen, M. (2011). *We Need a GitHub of Science*. Available at: <http://marciovm.com/i-want-a-github-of-science>
- Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received: 05 July 2011; paper pending published: 07 September 2011; accepted: 01 January 2012; published online: 24 January 2012.
- Citation: Lee C (2012) Open peer review by a selected-papers network. *Front. Comput. Neurosci.* 6:1. doi: 10.3389/fncom.2012.00001
- Copyright © 2012 Lee. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Maintaining live discussion in two-stage open peer review

Erik Sandewall<sup>1,2</sup>\*

<sup>1</sup> Department of Computer and Information Science, Linköping University, Linköping, Sweden

<sup>2</sup> Department of Publication Infrastructure, Royal Institute of Technology, Stockholm, Sweden

## Edited by:

Diana Deca, University of Amsterdam, Netherlands

## Reviewed by:

H. Steven Scholte, University of Amsterdam, Netherlands

Dietrich Samuel Schwarzkopf, Wellcome Trust Centre for Neuroimaging at UCL, UK  
Aliaksandr Birukou, European Alliance for Innovation, Italy

## \*Correspondence:

Erik Sandewall, Department of Computer and Information Science, Linköping University, S-58183 Linköping, Sweden  
e-mail: erisa@ida.liu.se

Open peer review has been proposed for a number of reasons, in particular, for increasing the transparency of the article selection process for a journal, and for obtaining a broader basis for feedback to the authors and for the acceptance decision. The review discussion may also in itself have a value for the research community. These goals rely on the existence of a lively review discussion, but several experiments with open-process peer review in recent years have encountered the problem of faltering review discussions. The present article addresses the question of how lively review discussion may be fostered by relating the experience of the journal *Electronic Transactions on Artificial Intelligence* (ETAI) which was an early experiment with open peer review. Factors influencing the discussion activity are identified. It is observed that it is more difficult to obtain lively discussion when the number of contributed articles increases, which implies difficulties for scaling up the open peer review model. Suggestions are made for how this difficulty may be overcome.

**Keywords:** open peer review, community peer review, two-stage peer review, live discussion

## 1. INTRODUCTION

Open peer review has been proposed for a number of reasons, in particular, for increasing the transparency of the article selection process for a journal, and for obtaining a broader basis both for feedback to the authors, and for the acceptance decision. It has also been proposed that the contents of the reviewers' comments and of the authors' responses to them may in themselves be of interest to the community of researchers in the area of the work, and that they should therefore be published and preserved.

Several of these goals rely on the existence of a lively review discussion. If the discussion falters then only the transparency goal remains, and if the discussion is limited to comments by two or three appointed referees and the authors' responses to them then the review process is little more than traditional peer review where merely the reviews are made publicly available.

Unfortunately, several experiments with open-process peer review in recent years have encountered the problem of faltering review discussions, for example, the experiment made by *Nature* in 2006 (Editorial Report: *Nature's Peer Review Trial*, 2006). It is therefore of interest to study examples of open peer review where it has been possible to maintain lively discussion, at least in some parts of the experiment, and to discuss the factors that may affect the volume and the character of the discussion.

The *Electronic Transactions on Artificial Intelligence* (ETAI) was an early experiment with the use of an open peer review process where lively review discussion was an explicit goal, and in fact an essential ingredient in the journal's review process. This journal was started by myself in 1997 because of my dissatisfaction with traditional peer review, and with an idea about an alternative peer review method that would not suffer from the same problems. Some parts of the journal's activities enjoyed lively review discussions; other parts did not. In this article I shall describe the experience from the ETAI in this respect and compare them with observations of one other two-stage peer review journal. I shall

observe that the problem of maintaining liveliness seems to be related to the question of scaling up of the journal's size, and conclude with suggestions for how scaling up may be achieved without sacrificing liveliness.

## 2. RATIONALE AND CONSTITUENCY FOR THE ETAI

Around years 1995 and 1996 I was concerned about the following problems with traditional, confidential peer review:

- The process can be manipulated. This is bad in itself, and it inspires distrust.
- If an article is rejected although its contents actually merit publication and this is discovered some years later, it is in practice impossible to correct the mistake and give due credit to the author. This is always damaging, and in particular so for articles that are ahead of their time.
- If an article is controversial, then the controversy should be brought out in the open so that everyone can make his or her own opinion about it. It should not be kept inside the close walls of the peer review process.
- Since reviewers are anonymous, they can not get proper credit for the work they put in. Quality control of the reviews is difficult for the same reason.
- Peer review is intended to serve two purposes: to provide feedback to the authors so as to improve the article, and to give a guarantee of quality. Its efficiency with respect to the first aspect is often marginal and could be improved.

Considerations similar to these have been discussed by many authors both before and after that time; see for example Gura (2002) and Benos et al. (2007). They led me to propose and to start the *Electronic Transactions on Artificial Intelligence* (ETAI)<sup>1</sup>

<sup>1</sup> <http://www.etaij.org/>

as an attempt toward the solution of these problems, without losing the strong points of conventional peer review. The research area being addressed by the ETAI is Artificial Intelligence, and some background about the character of this field is relevant for understanding the development of the ETAI itself.

Artificial Intelligence is a relatively independent branch of computer science that has strong connections to formal logic, formal linguistics, cognitive science, and a variety of other disciplines ranging from control engineering to psychology. The social structure of this partly interdisciplinary field of research is relevant for the ETAI peer review model: artificial intelligence can be viewed as consisting of a fairly large number of specialities, each with its own “college” of researchers that are active in the area, that meet regularly at conferences and workshops, and that to a large extent know about each others’ research directions. Each “college” has a worldwide membership that may count one or a few hundred researchers including the graduate students. The likely readers and the likely peer reviewers of a research article are usually found in the circuit of such a “college.”

Structures of this kind occur in many scientific disciplines but apparently not in all.

A second, important consideration concerns the character of research in the field. There is a combination of theoretical research and systems-building research. Theoretical research is done with standard methods of applied mathematics as applied to formal logic. Systems-building research is often done in large projects involving many participants over an extended period of time. It is generally acknowledged in the field that the results of systems-building research do not easily conform to the conventional publication formats, since it is difficult to identify “result modules” that are sufficiently independent of the rest of the large project and that can easily be published. Also, even if it is possible to construct a number of such “result modules” from a large project, the collection of these often fails to give a correct insight into the real results of the entire project. Finally, a large part of the real project results have such a character that they can best be communicated in a dialog-like setting where the pros and cons of different design decisions, for example, can be presented and discussed. They therefore do not fit so well into a framework where one expects to publish definite and unchallengeable results.

### 3. CONCEPTS AND DISTINCTIONS

The concept of “open peer review” is presently being used for several fairly different models of peer review. A basic distinction can be made between *open-names peer review* which is similar to traditional peer review except that the identity of the reviewers is shown openly, and *open-process peer review* where interested parties are invited to join the peer review process that takes place before an article is accepted for a journal or other similar venue. Hodkinson (2007) uses the term *community peer review* for open-process peer review and introduces additional distinctions.

One may notice that open-names and open-process approaches may be combined in several ways, so that one may use open-process peer review that does not operate with open-names, and vice versa. The present article will only address open-process peer review and will use the term *open peer review* as a synonym.

The ETAI used a *two-stage* peer review process (Sandewall, 1997b, 2006, 2009) that is based on both open-names and open-process, and that works in the following steps. Submitted articles are screened for relevance and if they pass this filter, they are posted on the journal’s webpage and made available to the community of researchers in the research area that the article addresses. This begins a 3-month period of open, constructive critique: questions are posed to the author, objections can be made and answered, and so forth. This *review* process is entirely open, so the names of all participants are seen openly (with rare exceptions). After the open discussion period, the author is able to revise the manuscript based on the feedback obtained, and resubmit it to the journal. It is then sent for *refereeing* to two or three referees whose identity is not divulged. The task of the referees is to only make a pass/fail decision, and they are asked not to propose additional changes in the article.

This separation of the peer review process into two stages reflects the two major purposes of peer review, namely, to improve the quality of submitted manuscripts, and to establish quality standard. Conventional peer review integrates these two goals, whereas in our system they are separated so that the purpose of the first stage is only for feedback to the author and for quality improvement, and the second stage is only for maintaining the quality standard. Therefore there is only one revision of an article, namely between the first and the second stage, and the version of the article that is submitted to the second stage becomes the final article if it is accepted. (This is the principle, but in fact there were occasional exceptions where a second round of minor revisions were requested).

The concept of *publication* needs to be made precise in the context of open-process peer review, in particular because of the very peculiar way that this word is used in the scientific community. The original and natural meaning of “publication,” in the sense of an activity, is of course to “make public.” However, in the context of scientific communication it is often considered to mean “published after having been accepted in a peer review process.” This terminology is problematic for us since open-process peer review requires by definition that articles are made available to the scientific community in its topic area for the purpose of *starting* the peer review process.

It is interesting to notice how this peculiar terminology has arisen in the first place. It can be led back to the establishment of the *Ingelfinger rule* (see, e.g., Angell and Kassirer, 1991), a principle developed by Franz Ingelfinger in the 1950s for use in the editorial offices of the New England Journal of Medicine, stating that this journal would not publish any articles whose contents were also published elsewhere, and requiring authors of submitted manuscripts to abide by this rule. The effect of this rule was to establish the journal as an archival one: if it is intended that annual volumes of a scientific journal are to be preserved in university libraries then it is inefficient to store several copies of the same article, whereas for journals that are received, read, and discarded this is not much of an issue.

The Ingelfinger rule was quickly adopted by many other journals at the time and has remained popular. Unfortunately however it was established only a few years before the spread of affordable small-scale reproduction technology using mimeograph

machines, and later on using large-volume copying machines. These had the effect that researchers in some fields started to prepare “departmental reports” for distribution to peers ahead of journal publication.

Journal editors and publishers reacted to this technical development in two different ways. In some areas, such as medicine, it was correctly observed that such departmental reports were publications, and according to the established rule the existence of such a report precluded publication of the same results in a journal, which of course effectively prevented the practice from being adopted at all. In other fields, such as mathematics, physics, and computer science, it was decided instead that departmental reports was a valuable thing to have, but instead of retracting the Ingelfinger rule one decided that a departmental report would not count as a *publication*, thereby making it possible for journals to accept such manuscripts. It is this game with the words that haunts us today.

This was an important issue when the ETAI was launched, in particular since one of the critical questions that we heard when we presented our novel peer review model was: *if an article is distributed openly before being accepted to the journal, how can one avoid that someone else “steals” the results and publishes them in his own name?* There was only one way of addressing this problem, namely, to return to the original meaning of “to publish” and to state as a terminological policy that an article was to be considered as published exactly when it was made public to the members of its peer community, which meant, well before it was accepted to the journal, and without any guarantee whatsoever that it would eventually be accepted. In this way, the priority for the results in the article should count from the date when the article was first made available.

This policy immediately led to a second question: if the article was published before being accepted for the journal, then who was the publisher? This led to the creation of the *Linköping University Electronic Press* (LiU E-Press)<sup>2</sup> as an open access publisher precisely for the purpose of having a publisher for submitted articles.

Consequently, whereas the Ingelfinger rule says that the journal will not publish previously published articles, our procedure implied that the journal would *only* publish previously published articles, namely, after the successful peer review of an article that had been published so that it could be peer reviewed.

These considerations concerning the concepts of publication and of publisher were laid out in an article that was published by the LiU E-Press in 1997 (Sandewall, 1997a). It was of course important to obtain as broad acceptance as possible of these unconventional ideas. I was therefore glad to have been invited to a working group that had been asked by the Association of STM Publishers (Science-Technology-Medicine) to find an answer to the question: *What should be considered as a publication in the electronic age?* – the problem being of course that there is no obvious original copy of a document that is produced and disseminated electronically.

The working group’s report (Frankel et al., 2000) reflects some of the ideas that have been described here, in particular insofar

as it recognizes several successive versions of a publication, where the peer reviewed version is designated as “final” but the earlier versions are also recognized as “publications.”

However, in my opinion the group never answered the basic question that had been posed, that is, how do you define the *publication* then? My own answer to this question was and is that one must first define an *electronic publisher* as an organization that is able to organize, preserve, and disseminate electronic documents persistently, and then define an *electronic publication* as an item that has been published by such an electronic publisher. The group did not however want to address this admittedly somewhat philosophical issue.

#### 4. CHALLENGES FOR A NEW PEER REVIEW MODEL

ETAI’s two-stage, open-process peer review model was easily accepted in its own research community of Artificial Intelligence. It was given particular strength since we secured the support of two important parties: it was published under the auspices of the Swedish Academy of Sciences and of the European Coordinating Committee for Artificial Intelligence, which is a federation of national A.I. societies.

This does not mean that everything was easy. The challenges were of several kinds:

- Doubts about the model by representatives of other disciplines, which in turn caused some of our colleagues to stay away from it.
- The problem of getting the flow of submissions to start initially.
- The problem of maintaining coherence in a journal that was divided between a number of specialized areas.
- Insufficiency of the computational and administrative infrastructure.

Any new journal of this kind is likely to face these questions, and it is important to be clear how a particular model for open peer review can handle them. I shall discuss them in turn.

##### 4.1. DOUBTS ABOUT THE OPEN-PROCESS PEER REVIEW MODEL

A number of persons told us that the ETAI peer review model simply would not work when it was first explained to them. Their pessimistic predictions turned out to be incorrect. It is interesting to note that the reason for the incorrect predictions was because people extrapolated from their acquaintance with traditional peer review but the extrapolation was not applicable.

In particular, one objection was that the model would not work since no one was going to contribute critical comments to the open peer review discussion for not risking to make enemies with the authors. This analysis was incorrect because whereas a critical comment in conventional peer review is to the author’s disadvantage (at least in an immediate sense), in the two-stage peer review scheme the author has a fair chance to respond to the critique, and also to make a correction in the article if this is warranted.

In fact, several of our authors reported that they *were glad* to receive critical comments since this made the discussion more lively, and therefore they obtained more attention for their article. This is like at a Ph.D. defense: a dull session is not appreciated, and the best is if the candidate obtains difficult questions and is able to answer them well.

<sup>2</sup><http://www.ep.liu.se/>



Another objection was that we would be overwhelmed by an avalanche of so-called “junk” articles, since authors would see a chance to have their articles published without peer review. This did not happen exactly because of the openness in the system. Under the conventional peer review scheme it does not “cost” anything to submit a substandard article since only the reviewers will know. In our model the quality of the article and the fact that it was not eventually accepted would be clear to everyone.

Predictions of this kind have appeared repeatedly, e.g., in an editorial of *Editorial: Revolutionizing Peer Review* (2005), but repeated practical experiences seem to refute it. The experience of the journal *Atmospheric Chemistry and Physics* (ACP) is similar to ours in this respect (Koop and Pöschl, 2006).

A complementary prediction was that we would not receive any submissions at all since no one would want to risk the shame of not having their article accepted. Fortunately it turned out that authors were more wise than that. We did decline some contributions and this did not have any noticeable effect on the flow of contributions afterward. Conversations with actual and would be authors suggested that this was not perceived as a problem.

Another objection concerned rejected articles. An article that has been rejected from a journal that uses conventional peer review can be submitted to another journal, but in our case this might not be possible, it was argued, since the article has been published in the formal sense. This did not seem to be a problem in practice, however, in particular since Computer Science is an area where prepublication using departmental reports is widely used and accepted, so journals tend to be generous in their interpretation of “previously published.” It might have been different in another field.

However, it should also be said that the practice where an author of a rejected paper resubmits the same paper to another journal without first acting on the reviewer feedback, is in fact a problem for the research publication system. Under the ETAI system it is still possible to submit repeatedly in this way, unless the second journal has a principle against it, but since the negative reviews from the ETAI are publicly available the author will have strong incentives to address the critique before the new submission.

Yet another objection was that the delay of 3 months until the acceptance of an article in the journal was too long. In the AAAS/UNESCO/ICSU workshop in 1998, Parker (1998) of the Royal Society of Chemistry stated<sup>3</sup>:

[This] contribution describes a very nice refinement to open review. However, I think most chemists would be horrified by the thought of peer review taking three months for the initial phase plus a bit longer for the intensive phase. The current average time from receipt to publication in RSC's flagship journal, *Chemical Communications*, is under 80 days and decreasing! I think this raises the distinct possibility of divergence of peer review policy among disciplines.

and later on:

Perhaps chemistry is less contentious and results less open to multiple interpretation than other disciplines. Certainly the

vast majority of decisions as to acceptance or rejection are very straightforward for chemistry articles using traditional peer review.

The observation that different disciplines operate under so different conditions that entirely different quality control schemes may be appropriate should of course be taken seriously. However, with respect to the time delay to “publication,” the question must be whether the chemists in this case want a quick decision in order to be able to disseminate the result to peers and obtain priority for it, or if it is in order to be able to put this additional merit item into his or her CV. If the former is the case then of course the delay time in the ETAI model is zero, since the result is disseminated and priority is established at the point where the review discussion starts. In the latter case, on the other hand, one will not be willing to accept substantial discussion periods, in particular if the character of the field is such that there is rarely much to discuss anyway.

In summary, we did have to work with explaining the two-stage open peer review model, and the important message had to be: in this system all the rules of the game are changed and all the habits change; you must think of it as an entirely different publication culture.

#### 4.2. STARTING THE FLOW OF SUBMISSIONS AND DEBATE

Another type of problem involved starting the entire process: not only getting the first submissions, but also getting the discussion to start for each of these. This was a chicken-and-egg situation: people were not likely to contribute to a discussion that no one listened to, but people would only listen if there were already some contributions.

The relatively unsuccessful experiment with community peer review in *Nature* in year 2006 (*Editorial Report: Nature's Peer Review Trial*, 2006) may possibly be due to this problem.

Under the ETAI system, the interested community for an article was notified using an email message when the article was presented for review discussion. This was maybe sufficient for getting some of these researchers to take a look at the article, but it did not suffice for getting the discussion started.

Two measures were instrumental for dealing with this problem in the ETAI. When the journal was entirely new, we presented its review scheme as having some of the features of a conference presentation, besides being a journal. At a conference you can present your work and get feedback on it, but in our journal you could have 3 months of discussion instead of 5 min, and the discussion was open to everyone in the research community in question and not merely those that attended the conference, and finally it was preserved and could be read (and continued) later on. As a continuation of the same parable we started panel discussions in the ETAI, where a few panelists made initial statements and then a discussion followed in our medium. This was effective in demonstrating to our constituency that if you send in a debate contribution then it is immediately seen by others, and this in turn encouraged submissions and debate contributions.

A second measure was taken if the discussion about a particular article did not start spontaneously: in those cases we could ask one or two colleagues to be discussion starters by making some initial comments. The experience was that once the discussion had started it tended to continue.

<sup>3</sup>www.aaas.org/spp/sfrl/projects/epub/ses3/parker2.htm

### 4.3. MAINTAINING COHERENCE

Our peer review model depended strongly on having an identifiable community whose members were likely to participate in the discussions. This was made possible by the fact that was mentioned initially, namely, that the research field of Artificial Intelligence is structured as a set of “virtual colleges” each having one or a few hundred members internationally. The mailing lists for the participants in these colleges were therefore essential for the functioning of the journal. Please recall that this was done long before the existence of social media; all communication had to be done using the journal’s website and communication by email.

The ETAI was therefore organized as a federation of specific research areas, each with its own area editor, its membership list, and so forth. Articles could only be submitted to a specific ETAI area and if there was no area that matched a particular article then it simply could not be submitted. Area editors were quite independent and operated their own wings of the journal.

The coherence and uniformity of the journal therefore became an issue. In retrospect I feel that I should have done more toward building the team spirit in the group of area editors; this would have made the journal stronger, it could have resulted in amore uniform appearance in the websites of the respective areas, and most importantly, it could have given help and support to the area editors in their work.

At the same time I do not think it would have been possible to work without the organization as a federation of areas. The task of the area editor in this scheme requires expertise and recognized standing in the area in question. It also demands much more work than being an area editor in a conventional journal, in particular because the area editor has to moderate the discussions about the submitted articles.

### 4.4. INSUFFICIENT COMPUTATIONAL AND ADMINISTRATIVE INFRASTRUCTURE

The publication and peer review scheme that was used by the ETAI required a computational infrastructure for the following purposes:

- For the publication of submitted articles, using the Linköping University Electronic Press.
- For the dissemination of information about newly submitted articles, and for the reception and dissemination of contributions to the discussion about an article. This was done using both email messages to the area members and additions to the area’s website.
- For the preparation of finally accepted articles in a form whereby the successive issues of the journal would have a graphic appearance that matched traditional journals.
- For the presentation of issues and volumes of the journal, containing both the actual articles and the review discussion for each of them.

These computational facilities were not ready when we started the journal; they had to be built as we went along. It would of course have been better to implement them first, but we had been eager to get started, we certainly underestimated the amount of work that was needed, and we did not know in advance what facilities

would be required. In any case, the requirement to build this software and, at the same time, to do the editorial work using partly improvised facilities led to a certain exhaustion on my part, and it was probably one of the factors that led to the discontinuation of the journal after a few years of relatively successful existence.

### 5. COMPARISON WITH CONVENTIONAL PEER REVIEW

An analysis of the advantages and disadvantages of a particular model for peer review should start with an identification of the goals that this process shall serve. Some such goals were mentioned in the Introduction, but there are in fact some additional goals that may be considered, as included in the following list.

- *Availability of reviewers:* insure that qualified reviewers will agree to participate and that they will wish to spend enough time and effort on the review assignment.
- *Amelioration:* improve the quality of a submitted article by providing feedback to the author.
- *Posterior use of reviews:* are the reviews valuable after the end of the peer review period?
- *Selection:* acceptance to the journal confirms that the article meets a specific quality standard, which helps readers decide which articles to read.
- *Fairness:* it is not merely in the interest of the readers, but also in the interest of the authors that acceptance decisions are fair and unbiased.
- *Merit:* acceptance of the article contributes to the author’s scientific credentials.
- *Attention:* in the case of open-process peer review, the discussion in that process gives attention to the article in the researcher community of the article’s topic.

We shall use this list as a framework for comparing the ETAI model for two-stage peer review with the conventional, blind review model.

The Attention aspect is by definition not present for conventional peer review. Authors in the ETAI reported that for them it was an important and positive aspect of the review model.

Conventional peer review integrates the Amelioration and Selection aspects into one single process. In two-stage peer review the two stages are dedicated to the Amelioration goal and the Selection goal, respectively.

The quality of the process with respect to Amelioration and Selection depends of course entirely on the competence and the efforts of the reviewers. I can only provide a subjective and qualitative estimate of this, based on also having been co-Editor-in-Chief of the journal Artificial Intelligence, AIJ (the most prestigious journal in its area) for a number of years, besides of course my general experience of other journals. My experience is that the quality of reviews varies enormously between journals, and that the quality of reviews (i.e., contributions to the open review discussion) in the ETAI was in the upper-middle range. It could not match the AIJ, but it was as good or better than many others.

One way of estimating the Selection effect is to check the acceptance rate, with an assumption that a low acceptance rate in a journal indicates that only articles with very high quality will be accepted there. In the case of the ETAI, the number of declined

articles was quite low. This might be an indication to its disadvantage, but there are some considerations that should also be taken into account. First of all, the numbers may not be comparable due to the “shame” effect that was discussed above: it is likely that authors thought carefully before submitting an article, in consideration of the risk of having it declined, and if this is true then the overall quality of submitted articles would tend to be higher. I have no way of quantifying this, but the argument suggests that one should be careful when comparing acceptance rates for the two peer review systems.

Another question in this context is whether it is truly in the interest of the scientific community that a journal is very restrictive with acceptances? For example, if reviewers have widely different assessments of an article and neither reviewer is willing to change their opinion, is this then a reason for accepting the article or for rejecting it? A strong emphasis on “quality” implies a reject decision, but this may effectively stop new and truly important contributions.

The usual argument in favor of a strict acceptance policy refers to the Selection goal: readers have limited time at their disposal, and the peer review process shall assist them by filtering out the articles that are required reading. Notice, however, that this is one more example of how the analysis departs from the characteristics of the conventional peer review system, without taking the effects of the alternative system into account. This is because in the conventional system, the *only* information that is available to the reader for his or her selection decision is the binary information that the article was accepted, plus of course the information about and by the author, such as the abstract. In the open-process model, on the other hand, the would be reader may check the *discussion* about the article as a first indication of whether the article is worth reading or not for him.

In general, the more meta-information you have about an article, the better. The abstract and the record of the discussion play different and complementary roles. As a reader, the information about the author and the author’s institution gives some cues about quality and relevance. The title and the abstract are important for identifying whether the topic is relevant for him. The record of the discussion moderates these first impressions with respect to both quality, relevance and novelty. Consequently, a journal with open-process peer review may be somewhat more generous with its acceptances, thereby reducing the risk of missing important original developments, and still provide its readers with enough information that they can select their reading menu efficiently.

Another argument with respect to acceptance policies is that the acceptance of a marginal article tends to reduce the journal’s impact factor. The argument goes as follows. It is known that the distribution of citation counts is extremely skewed, so that a small number of articles obtain very many citations, and most articles obtain few. Since the impact factor for a journal is calculated as the arithmetic average of the citation counts for all articles in the journal, any article whose citation count is lower than the journal’s average will reduce its impact factor. Moreover, although one must be sympathetic to the problems of getting groundbreaking articles published, the hard fact is that they will only gain attention after a number of years, whereas impact factors are calculated based on citation counts during only a few years after publication.

Therefore, publication of such (rare) articles does not contribute favorably to the journal’s impact factor.

The only thing one can say about this argument is that it illustrates the irrational character of the use of impact factors, and its detrimental effects on the scientific publication system.

The goal of Fairness is an important one. Benos et al. (2007) expressed doubt that open-process peer review would represent an improvement in this respect; they wrote:

Both of these journals (ACP and ETAI) do not unmask the people who decide whether or not a paper is publication worthy. . . . This does not remove any bias, perceived or real, by referees or editors. Thus, these forms of open review, while alleviating the delays and increasing transparency, will not attenuate perceptions of bias at the actual acceptance step of the process.

This analysis is incorrect, for two reasons. First, the transparency of the review discussion and the attention that it provides for the article before the acceptance decision is a significant safeguard against malpractice in the refereeing stage. Secondly, even if an article is declined in the refereeing stage in the ETAI, it will still have the advantage of first publication with the ensuing citability and the proof of priority of the results. This means that a mistaken decision to decline or reject an article, should it occur, is much less detrimental for the authors and the article than what it is when the conventional peer review process is used.

A final remark concerns the Merit aspect of the peer review process. One consequence of the rapid growth of science and of scientific publication is that researchers and research projects are increasingly evaluated based on numbers that represent their publication and citation scores, whereas in older times it was taken for granted that in order to evaluate a person’s research you must read and evaluate his or her publications. There are many voices to the effect that the numerical evaluation is very unsatisfactory, but the argument is anyway that we do not have any choice, in view of not only the amount of reading that would otherwise be required, but also the increasing specialization whereby reviewers are frequently called on to assess and to compare candidates whose area of research they do not themselves master. The persistent availability of the review discussion for an article may alleviate this problem, since even an outsider may often get a good notion of a researcher’s standing and the quality of her work by hearing or reading an exchange of opinion between this person and his or her peers.

This possibility requires however that the discussion about each article is sufficiently extensive, which again adds to the reasons why it is in the interest of an author to have as many contributions to his review discussion as possible, including in particular a number of critical contributions that it is a challenge to answer.

## 6. MAINTAINING LIVELINESS IN PEER REVIEW DISCUSSIONS

As one can see from the ETAI webpage, some parts of the journal enjoyed lively peer review discussions, and in other parts the discussion did not really get off the ground. As stated in the Introduction, it is of great interest to understand the factors behind this difference.

## 6.1. PAST EXPERIENCE

Almost the first things that we learnt after starting the ETAI was that discussions do not usually start by themselves. Merely posting articles on the journal's website and inviting contributions is not very effective. I have described the methods that we used for starting discussions, and some of the cases of failed discussions may have been due to the insufficient use of these methods.

However, looking in retrospect at the ETAI experience it seems that another factor was also important, namely the question of *reader fatigue* and the related question of *limited exposure*. In those cases where a reader of the journal was exposed with a considerable number of articles in the same short period of time, it seemed that it was difficult to get the reader to engage herself or himself in any of these articles, whereas if only a few articles were offered and these were quite relevant to his interests, then it was much more likely that he or she would write a debate contribution. The partitioning of the journal and the readership into areas of limited size insured that each reader of the journal received a sufficiently limited exposure and a sufficiently focused set of new articles per time unit for her or his consideration.

The hypothesis that a limited reader exposure was important for insuring good participation in the discussions is not something that we can validate by hard data; it is only based on a general understanding of how our readers operated. It is however consistent with the actual discussion intensity in the ETAI, and in particular with the outcome of our attempts to base special ETAI "sections" on contributions at specialized workshops. The idea for this was simple: such a workshop engages the same "virtual college" as is used for defining an Area within the ETAI, workshops are used both for presentation of recent work and for discussions, and the ETAI seemed to be a natural way of extending both those aspects of the workshop activity. To begin with, we would invite the workshop participants to write down their main comments at the workshop and to contribute them to the ETAI.

This worked very well in one case, and not very well in several others. The *Special Section on Knowledge and Reasoning in Practical Dialog Systems*<sup>4</sup> is a case where it worked very well, but it also required a considerable effort by the area editors for obtaining and editing the debate contributions from the workshop participants. On the other hand, when individual articles were submitted one at a time it was easier for an area editor to obtain a viable discussion.

It is interesting to compare this experience with the situation in the journal *Atmospheric Chemistry and Physics* (ACP; Koop and Pöschl, 2006)<sup>5</sup> which is arguably the most successful example of two-stage open-process peer review at present, and which started its operation in 2001. The peer review procedure in the ACP, as described on its website, is in principle quite similar to the one used by the ETAI, but with one major difference: the ETAI was organized as a federation of areas and the discussion was primarily viewed as an internal discussion within each area, but the ACP does not have such a structure. All submitted articles are presented in a single, chronological list on the ACP webpage, and the interested reader

will see all of them. Furthermore, the publication volume of the ACP is significantly higher than for the ETAI.

It is against this background that one must read the statistics about the participation in review discussions in the ACP. For example, as observed on May 15, 2011, among the 41 submissions that had been received between March 1 and March 15, 32 had obtained no or one contribution to the discussion. Six of them had obtained 2 contributions, and 3 of them had obtained 4 contributions. Among the 24 discussion contributions in the discussions with more than one contribution, only 5 were by third-party persons and the other 19 were by a designated referee or by the authors. These figures apply 2 months or more after the beginning of the discussion. For the 39 articles received between May 1 and May 15, only one of them had even one discussion contribution.

It seems, therefore, that although the ACP is a very impressive example of the use of open-process peer review, the most important aspect of its model is that it advances the transparency of the review process, and that it guarantees that articles are published and citable from the very beginning of that process. On the other hand, if one is interested in obtaining a real community discussion about submitted articles, then the ACP does not offer a strong case.

As already mentioned, the approach used by the ETAI was relatively labor intensive for each of the area editors, and it only covered some parts of Artificial Intelligence. Consider, therefore, the question how one would organize a journal that used open-process peer review with lively discussions and that was anyway able to publish several hundred articles per year. How would it be organized, given what has been said about the need to both encourage and to moderate the discussions about each article. This is the question that must be answered if the strong aspects of the ETAI experiment is going to scale up.

## 6.2. A FIRST PROPOSAL

The first step toward answering this question must be to obtain a clear understanding of the structure of the scientific discipline that the journal would serve. Does it resemble the structure of Artificial Intelligence where there are identifiable specialties with their own problem statements, memberships, workshops, cooperations, and competitions, and is the difference only that the number of such specialties is much larger? Alternatively, does it instead have a more open structure where researchers continuously monitor research articles and results that emanate from a much larger population of fellow scientists?

In the former case I imagine that it should be possible to scale up the approach that was used by the ETAI while using the Wikipedia organization as a model. Concretely speaking, it would be necessary to organize the resulting large number of areas and area editors using a firm set of rules and guidelines for all aspects of the journal's operation, and to have a reliable and complete computational infrastructure already from the start of the operation. These were things that the ETAI did not have.

## 6.3. A SECOND PROPOSAL

In the latter case, it seems clear that the ETAI model would not work: having a large number of members in an area for the journal

<sup>4</sup><http://www.ida.liu.se/ext/epa/ej/etai/1999/D/>

<sup>5</sup><http://www.atmos-chem-phys.net/>

would put an unreasonable workload on the area editor, and our informal observations of the importance of reader fatigue suggests that participation in the discussions would anyway be too low. Moreover, the observations of actual debate participation in the ACP suggests that its model will also not be able to support lively discussions.

I will therefore offer the following proposal for how to organize a larger journal in this case: one may try using a system based on *ad-hoc discussion groups*. For each article, or for a small set of related articles, one would form a discussion group that should last for the entire review period of the article(s) in question. Peers should not be enabled to make discussion contributions randomly in the full set of articles that are under discussion, but only by joining a discussion group and staying with it. In order to insure continuity and coherence in the discussion, a participant in the journal's discussion activities could be encouraged to engage in a reasonable number of groups at each point in time, and to join a new group when one that she is in has completed its work, i.e., the acceptance decision has been made. The identification of a new group to engage in could be made through invitation by another group member ("Here's an article that you'd find interesting") or by active search by the participant, or by a service where the software system suggested relevant groups.

An important consideration would then be to strive for a good mix of participants in each ad-hoc group, in particular, to engage the entire range from Ph.D. students to senior researchers. In fact, an advisor might find it worthwhile to *require* her or his students to participate in a number of such groups as one part of their Ph.D. study.

The purpose of organizing such ad-hoc discussion groups would be to arrange a level of contact between reader and journal where limited and focused reader exposure is obtained, and where it should be possible to attract and retain the reader's attention to a limited number of articles. An obvious problem with this model would be that some articles may attract a very large number of discussants, and others may not attract any. The former problem should not be handled by creating several groups, since it would overburden the author; it would be better to simply let the system enforce a limit on the number of discussants for each article. The problem of no discussants or too few discussants is more difficult, but one possibility would be to refer such articles to conventional peer review.

Another possibility would be to decide that if no one is interested then the article is automatically declined for the journal. Such a policy would not be as harsh as it may sound, since the likely effect of it would be that each author would try to engage a certain number of discussants for her or his article. Hopefully this would be sufficient for avoiding the situation where a perfect paper is dropped because no one has anything critical to say about it. The scheme might however bias the discussion in a too positive and uncritical direction. This can only be determined by actually experimenting with this policy as well as alternative ones.

## 7. ADDITIONAL ASPECTS OF TWO-STAGE PEER REVIEW

Although the question of maintaining liveliness of discussion even in the case of scaling up is the most important issue, there are anyway some other aspects of two-stage open-process peer review

that may be discussed in the light of the experiences that have been described.

### 7.1. SHOULD OPEN-PROCESS PEER REVIEW USE AN OPEN-NAMES POLICY?

With the experience from having operated the ETAI it is interesting to read about other experiments with open-process peer review as well as reading more general comments and proposals in the same direction. It is striking that many of them make the same extrapolations from the culture of conventional peer review as we encountered when the ETAI was started. In particular, it is frequently argued that the identity of the discussants must be kept confidential because otherwise the comments will be very dull; see e.g., Suls and Martin, (2009), or Khan (2010) for an editorial in the British Medical Journal. Our experience was however contrary to observations such as these, for the reasons that were stated above.

There was in fact one particular occasion when a discussant requested that his name should be withheld, but for an interesting reason: he had made similar, critical remarks to the same article when it had previously been submitted to a conventional journal, and rejected, and if his name were to be stated in the ETAI discussion then he feared that the author would be tired because of the role he had played in the decision of that other journal. This illustrates how it is the character of the conventional peer review process that *causes* reviewer anonymity to be an issue, and not the phenomenon of critique in itself.

To the extent that lively review discussion is considered as an important goal, so that transparency of the review process is not the only consideration, it is also plausible that an open-names policy with respect to all participants in the discussion will increase the attention that is paid to the discussion, and therefore, will tend to increase the number of further contributions to it. Knowing who has written a contribution to a discussion adds to the reader's perspective on it and is likely to stimulate her or his opinions on the matter. It follows also that an additional advantage of the open-names policy is that it may help strengthening the community of researchers in question, and in particular to help including those that are not able to travel to the important conferences.

### 7.2. DURATION OF THE COMMENTARY PERIOD

Several proposals for open peer review suggest that the discussion should go on for an unlimited time, and in some cases that there should not be any strict acceptance decision but merely an initial screening for relevance and appropriateness of a submitted article. This means in effect that only the first stage of the ETAI two-stage process is used, and it goes on indefinitely. However, even in the two-stage process there is absolutely no reason why one should not be able to add further comments to the discussion after an article has been accepted, or after it has been declined, and in the latter case this might also lead to the article being reconsidered for acceptance<sup>6</sup>. On the other hand I still believe that there is a value in having a limited period of time when particular attention is given

<sup>6</sup>This indicates in fact an additional advantage of open-process peer review: if an article has been declined mistakenly then the mistake can be corrected later on and the author can receive due credit. In the conventional peer review system it is very difficult to correct such mistakes.



to the article, so that one can obtain a coherent discussion about it and not merely a number of occasional comments.

The question of what is the optimal duration of the commentary period is an important one. If it is too short then it will not give peers enough time to think and to react; if it is too long then peers may be led to postpone making their contributions, which leads to a loss of dynamism in the discussion. Moreover, the observation concerning reader fatigue suggests that commentary periods should be kept short, so that the set of articles under discussion at any one time is kept fairly small. Different journals and different disciplines may strike this balance in different ways. In the case of the ETAI I think the 3-month period was reasonable, but 2 months would probably also have worked well.

### 7.3. ARTICLE PUBLICATION STATUS DURING THE REVIEW PHASE

An additional difference between the peer review procedures in the ETAI and the ACP concerns the publication of articles at the beginning of the review debate. In the design of the ETAI procedure we were very concerned about the publication status of a submitted article during its discussion period, and as explained above we defined a mechanism whereby the article would count as *published* on the date when it was advertised and made available to its peer community for the purpose of discussion, in particular so that it would count for priority of results. We created the Linköping University Electronic Press for this purpose, and we participated in the discussion at that time about what constitutes an electronic “publication.”

The ACP has chosen another approach: concurrently with the ACP journal there is the journal-like *Atmospheric Chemistry and Physics Discussions* (ACPD) whose webpage is graphically similar to its parent journal, but where it is made clear that articles are included there prior to peer review and eventual acceptance in the ACP.

The approach used by the ETAI was more elaborate. We chose it because of a long-term consideration where we wanted research articles to be associated with research data and with computational processes that illustrate and validate the contents of the articles themselves. Such attachments to articles impose particular demands with respect to long-term maintenance, and it was not possible to make such guarantees in our E-Press for all ETAI authors that might wish to use such facilities. Instead, the strategy was to encourage other institutions in our area to set up their own counterparts of the E-Press, so that both the pre-review publication of the article itself and the definite publication of the attached resources should be done in the author’s home institution, or in an entity that was dedicated to this service – a kind of “web hotel” for research articles and their related materials.

It turned out that no other institution reacted to this suggestion during ETAI’s active period, so in practice the Linköping E-Press ended up doing the initial publication of all submitted articles, as well as of course the ETAI journal itself. However, I still believe that the proper organization of attached computational materials is an important issue for the future, at least for our field of research and probably for many others.

Another consideration with respect to publication status and priority arises with respect to how we defined the date of publication of an article. Since we considered in principle that the

starting date of the discussion period was the date of publication of the result, we used it for defining the date of publication of the final article. Thus an article whose discussion started in October of year X and that was accepted for the journal in February of year X + 1 would appear in the journal issue for October–December of year X. The logic behind this was clear, but it was not always easy to explain it to authors and readers.

This design led in turn to another consideration, namely, a restriction on what changes were permitted in an article between the original submission and the final version for the journal. On one hand we wished of course that the review discussion should result in improvements, but on the other hand it would have been unfair if the final version were to contain essential results that had been obtained after the publication (in our sense) of the first version. There was a rule, therefore, that the changes should be restricted to improvement of presentation, without strengthening the results as such.

In one concrete case, an author of a relatively theoretical article reported during the discussion period that he had some additional results that would fit well into the same article, and the question was what to do with them. The solution was that his additional results were written up as “short note” that was presented as an addition to the original article, but with a later date of publication. Such a separation of the results would have been inconvenient in a paper-based journal, but in the electronic medium it was not a big issue.

These considerations with respect to publication date may seem unnecessary, but my view on this is that they should be viewed in the same way as formal business contracts in one’s personal life: as long as the relations between people are dominated by common sense there is no need for formality, but if problems should arise then they can be handled with less pain if there are clear rules and clear data. Priority of research results is sometimes a topic of considerably animosity, and it is worthwhile to design one’s publication system in such a way that one has a firm basis for resolving conflicts at those rare occasions when they do arise.

### 7.4. INNOVATIVE SOFTWARE TECHNIQUES VS. CLASSICAL STYLE

Several of the measures that we took in order to make the ETAI acceptable are no longer needed, and may be irrelevant for future introduction of two-stage peer review. We organized our journal in terms of annual volumes and issues, with consecutive page numbering throughout each volume, although in principle it would have been more natural to consider an annual volume just as a set of articles and to number the pages of each article from one and up. We also produced a small supply of paper-printed copies of each issue, with a nice-looking cover, so that we could show it at conferences and archive it in major libraries. Measures such as these are superfluous today, or will soon be.

The computational infrastructure that was used by the ETAI seems antiquated by contemporary standards. Today we would certainly use a more interactive implementation. It would be natural to consider using wiki techniques and social-media techniques.

At the same time I would be careful not to go overboard with the use of modern software paradigms. For good and for bad, prestige is an important factor for a scientific journal, which means it must

inspire confidence and signal continuity. This applies not only for the articles that are submitted, debated and eventually accepted, but it applies as well for the discussion. In the case of the ETAI we made sure that the discussion contributions were presented in a correct fashion. In fact, one of the ETAI areas actually operated a side-journal called an Electronic Newsletter that was dedicated to presenting the discussion contributions, as well as other information of interest, in a nicely formatted form that resembled the format of the main journal. This was done in order to give prestige, in a good sense, to the discussion contributions so that people should feel that these discussions were valuable material: valuable to read, and valuable to have written, something that you could add to your C.V.

One other aspect of the prestige policy was to maintain a high conversational standard in the review discussions, besides of course a high scientific standard. The discussion was moderated, no contribution appeared on the website until it had been approved by the area editor, and the tone of critical comments was monitored. In fact, it is not so uncommon that reviewers in conventional peer review take advantage of their anonymity for adopting a condescending tone vis-a-vis the author and the submitted article. Some discussants retained the same haughty attitude in their contributions to our discussion. We therefore imposed a strict policy of asking the discussant in such cases to revise the wording and to adopt a tone that he would use if he talked to the author face to face and in a civil manner.

My suggestion for a contemporary open-process peer review scheme would therefore be to carefully consider all that can be offered by modern Internet-related technology, but to only adopt it when it is compatible with a policy of consistently good style and effective quality control of all aspects of the journal's operation.

## 7.5. BEYOND CONVENTIONAL ARTICLES: PEER REVIEW IN NEW ENVIRONMENTS

Innovation in the publication and communication of research results is not confined to the well-known topics of electronic publishing and open access, or to the current topic of changing the peer review model. The present article has discussed alternative peer review but with an assumption that the character of the articles themselves has not changed. This assumption will not remain valid for long. There is an abundance of new topics when other kinds of publications are considered, and here I can merely indicate my own particular interests in this respect. One important topic concerns the organization of *evolving articles* where the author of an accepted article is made responsible for the update and maintenance of the article during a period of time and is able to amend it successively (Sandewall, 2010). I am also interested in the question of publication of information modules whose contents range from "facts" to "knowledge," and how such modules can be published, peer reviewed, cited, and so forth (Sandewall, 2008, see also the Common Knowledge Library<sup>7</sup>). Finally there is an interesting issue concerning how to organize a publication mechanism that is appropriate for publishing the results of large, integrated, systems-oriented projects. All these new kinds of publications will require novel forms of peer review that are adapted to

their peculiar characteristics. I am convinced that an open-process peer review scheme will be appropriate in those cases as well, but the basic setup will be different from what you need for peer review of conventional articles.

## 7.6. COEXISTENCE BETWEEN PEER REVIEW SCHEMES

One of the most important observations from the ETAI experiment is that open-process peer review creates and requires a culture that differs from conventional peer review in important ways. The change of rules and practices affects the expectations and the behaviors of authors and of reviewers in ways whereby these behaviors tend to gravitate to a new and different equilibrium, so to say.

This raises the question as to what will happen when conventional and alternative methods of peer review coexist. Several scenarios are possible. One may imagine a polarization where some research communities embrace the new methods wholeheartedly and other communities reject them outright. One may also imagine the emergence of intermediate models: a kind of "open peer review light." Finally one may imagine a kind of "survival of the fittest" in the competitive world of research publication, namely, if the disadvantages of belonging to the minority that uses a non-standard scheme are so big that it can not survive in the long run. For example, quantitative research assessment constructs such as impact factors and acceptance rates are based in the culture of conventional peer review, and furthermore they tend to favor existing journals over new ones. If they are applied to publication venues that use alternative peer review schemes then these may easily find themselves at a disadvantage in several ways.

## 8. CONCLUSION

In this article I have discussed the experience from the Electronic Transactions on Artificial Intelligence and made some suggestions for what would be needed in order to scale up the size of a journal with open-process peer review without sacrificing the liveliness of the review discussion. An additional theme of the article has been that the use of the combination of open-names and open-process, two-stage peer review tends to change the researchers' perceptions and expectations in the review process in a multitude of ways, and that it can easily be very misleading to try to predict what will happen in such a scheme by extrapolation from what is the case when conventional peer review is used.

This observation is in opposition to a suggestion made by Stevan Harnad when he wrote as follows (Harnad, 1997):

Peer review is imperfect; it can no doubt be improved upon, but alternatives should first be tested; and in testing, one is well-advised to manipulate one variable at a time: Here we are dealing with a change in medium (paper to electronic), a change in economic model (subscription to author-side payment) and a change in quality control mechanism (peer review to open peer commentary).

As we have seen there is a number of other "variables" that are also being changed, and the problem is that the effects of those changes are not independent. There are clear indications that when a change of one variable at a time is likely to have one set of consequences, the effects of changing several of them together may have

<sup>7</sup><http://piex.publ.kth.se/ckl/>

consequences that are quite different from the individual changes. This is a reason why the topic of alternative methods for peer review is so difficult to analyze, and such a fascinating challenge to experiment with.

## REFERENCES

- Angell, M., and Kassirer, J. P. (1991). The Ingelfinger rule revisited. *N. Engl. J. Med.* 325, 1371–1373.
- Benos, D. J., Bashari, E., Chaves, J. M., Gagar, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., McGowan, S., Polter, A., Qadri, Y., Sarfare, S., Schultz, K., Splittgerber, R., Stephenson, J., Tower, C., Walton, R. G., and Zotov, A. (2007). The ups and downs of peer review. *Adv. Physiol. Educ.* 31, 145–152.
- Editorial: Revolutionizing Peer Review? (2005). *Nat. Neurosci.* 8, 397.
- Editorial Report: Nature's Peer Review Trial (2006). *Nature*. doi: 10.1038/nature05535
- Frankel, M. S., Elliott, R., Blume, M., Bourgois, J.-M., Hugenholtz, B., Lindquist, M. G., Morris, S., and Sandewall, E. (2000). Defining and certifying electronic publication in science. *Learn. Publ.* 13, 251–258.
- Gura, T. (2002). Scientific publishing: peer review, unmasked. *Nature* 416, 258–260.
- Harnad, S. (1997). *Listserv Comment on 'Open Peer Commentary: A Supplement, Not a Substitute, for Peer Review.'* Available at: <http://list.uvm.edu/cgi-bin/wa?A2=ind9706&L=serialst&D=0&P=4500&F=P>
- Hodkinson, M. (2007). Open peer review and community peer review. *Journalogy*. Available at: <http://journalogy.blogspot.com/2007/06/open-peer-review-community-peer-review.html>
- Khan, K. (2010). Is open peer review the fairest system? No. *Br. Med. J.* 341, c6425.
- Koop, T., and Pöschl, U. (2006). An open, two-stage peer review journal. *Nature*. doi:10.1038/nature04988
- Parker, R. (1998). "Response to Sandewall's alternative view to peer review," in *AAAS/UNESCO/ICSU Workshop on Developing Practices and Standards for Electronic Publishing in Science*, Paris.
- Sandewall, E. (1997a). A neo-classical structure for scientific publication and reviewing. *Linköping Electronic Articles on Academic Policies and Trends*. 2, nr 1.
- Sandewall, E. (1997b). Publishing and reviewing in the ETAI. *ETAI* 1, 1–12.
- Sandewall, E. (2006). Opening of the process. A hybrid system of peer review. *Nature*. doi:10.1038/nature04994
- Sandewall, E. (2008). Extending the concept of publication: factbases and knowledgebases. *Learn. Publ.* 2, 123–131.
- Sandewall, E. (2009). "Experience of two-stage peer review in the ETAI, 1997–2001," in *International Symposium on Peer Review*, Orlando.
- Sandewall, E. (2010). Exercising moral copyrights for evolving publications. *ScieCom Info* 6, 1–4.
- Suls, J., and Martin, R. (2009). The air we breathe. A critical look at the practices and alternatives in the peer-review process. *Perspect. Psychol. Sci.* 4, 40–50.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 15 July 2011; paper pending published: 07 August 2011; accepted: 03 February 2012; published online: 21 February 2012.

Citation: Sandewall E (2012) Maintaining live discussion in two-stage open peer review. *Front. Comput. Neurosci.* 6:9. doi: 10.3389/fncom.2012.00009

Copyright © 2012 Sandewall. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Tracking replicability as a method of post-publication open evaluation

Joshua K. Hartshorne\* and Adena Schachner

Department of Psychology, Harvard University, Cambridge, MA, USA

## Edited by:

Nikolaus Kriegeskorte, Medical  
Research Council Cognition and Brain  
Sciences Unit, UK

## Reviewed by:

Nikolaus Kriegeskorte, Medical  
Research Council Cognition and Brain  
Sciences Unit, UK  
Alexander Walther, Medical Research  
Council Cognition and Brain Sciences  
Unit, UK

## \*Correspondence:

Joshua K. Hartshorne, Department of  
Psychology, Harvard University, 33  
Kirkland Street, Cambridge, MA  
02138, USA.  
e-mail: jharts@wjh.harvard.edu

Recent reports have suggested that many published results are unreliable. To increase the reliability and accuracy of published papers, multiple changes have been proposed, such as changes in statistical methods. We support such reforms. However, we believe that the incentive structure of scientific publishing must change for such reforms to be successful. Under the current system, the quality of individual scientists is judged on the basis of their number of publications and citations, with journals similarly judged via numbers of citations. Neither of these measures takes into account the replicability of the published findings, as false or controversial results are often particularly widely cited. We propose tracking replications as a means of post-publication evaluation, both to help researchers identify reliable findings and to incentivize the publication of reliable results. Tracking replications requires a database linking published studies that replicate one another. As any such database is limited by the number of replication attempts published, we propose establishing an open-access journal dedicated to publishing replication attempts. Data quality of both the database and the affiliated journal would be ensured through a combination of crowd-sourcing and peer review. As reports in the database are aggregated, ultimately it will be possible to calculate replicability scores, which may be used alongside citation counts to evaluate the quality of work published in individual journals. In this paper, we lay out a detailed description of how this system could be implemented, including mechanisms for compiling the information, ensuring data quality, and incentivizing the research community to participate.

**Keywords:** replication, replicability, post-publication evaluation, open evaluation

## IMPROVING THE QUALITY OF PUBLISHED RESEARCH

The current system of conducting, reviewing, and publishing scientific findings – while enormously successful – is by no means perfect. Peer review, the primary vetting procedure for publication, is often slow, contentious, and uneven (Mahoney, 1977; Cole et al., 1981; Peters and Ceci, 1982; Eysenck and Eysenck, 1992; Newton, 2010). Incorrect use of inferential statistics leads to publication of spurious findings (Saxe et al., 2006; Baayen et al., 2008; Jaeger, 2008; Kriegeskorte et al., 2009; Vul et al., 2009; Wagenmakers et al., 2011). Publication biases, such as the bias against publishing null results (e.g., Easterbrook et al., 1991; Ioannidis, 2005b; Boffetta et al., 2008), lead to distortions in the published record, hampering both informal reviews and formal meta-analyses. Numerous valuable proposals have been offered as to how to improve the system in order to enable researchers to better identify high-quality research, including those in the present special issue.

There are many considerations that go into determining research quality, but perhaps the most fundamental is replicability. Recently, numerous reports have suggested that many published results across a range of scientific disciplines do not replicate (Ioannidis et al., 2001; Jennions and Möller, 2002b; Lohmueller et al., 2003; Ioannidis, 2005a; Boffetta et al., 2008; Ferguson and Kilburn, 2010). However, because replication attempts are not tracked and are often not reported, there is no systematic way for researchers to know which results in the literature have been replicated.

In the present paper, we first discuss evidence that the rate of replicability of published studies is low, including novel data from a survey of researchers in psychology and related fields. We propose that this low replicability stems from the current incentive structure, in which replicability is not systematically considered in measuring paper, researcher, and journal quality. As a result, the current incentive structure rewards the publication of non-replicable findings, complicating the adoption of needed reforms. Thus, we outline a proposal for tracking replications as a form of post-publication evaluation, and using these evaluations to calculate a metric of replicability. In doing so, we aim not only to enable researchers to easily find and identify reliable results, but also to improve the incentive structure of the current system of scientific publishing, leading to widespread improvements in scientific practice and increased replicability of published work.

## WHY MIGHT WE EXPECT LOW REPLICABILITY?

Many aspects of current accepted practice in psychology, neuroscience, and other fields necessarily decrease replicability. Some of the most common issues include a lack of documentation of null findings; a tendency to conduct low-powered studies; failure to account for multiple comparisons; data-peeking (with continuation of data collection contingent on current significance level); and a publication bias in favor of surprising (“newsworthy”) results.

## LACK OF PUBLICATION OR DOCUMENTATION OF NULL FINDINGS

Null results are less likely to be published than statistically significant findings. This has been extensively documented in the medical literature (Dickersin et al., 1987, 1992; Easterbrook et al., 1991; Callaham et al., 1998; Misakian and Bero, 1998; Olson et al., 2002; Dwan et al., 2008; Sena et al., 2010), with additional reports in political science (Gerberg et al., 2001), ecology and evolution (Jennions and Möller, 2002a), and clinical psychology (Coursol and Wagner, 1986; Cuijpers et al., 2010). There appear to be fewer comprehensive studies of publication bias in non-clinical psychology, although evidence of this bias has been documented in a few specific literatures (Field et al., 2009; Ferguson and Kilburn, 2010).

Preferential publication of significant effects necessarily biases the record. Consider cases in which multiple labs all test the same question, or in which the same lab repeatedly tests the same question while iteratively refining the method. By chance alone, some of the experiments will result in publishable statistically significant effects; the likelihood that a finding may be spurious is masked by the fact that the null results are not published.

The significance-bias also leads to the overestimation of real effects. Measurement is probabilistic: the measured effect size in a given experiment is a function of the true effect size plus some random error. In some experiments, the measured effect will be larger than the true effect, and in some it will be smaller. Suppose the statistical power of the experiment is 0.8 (a particularly high level of power for studies in psychology; see below). This means that the effect will be statistically significant only if it is in the top 80% of its sampling distribution. Twenty percent of the time, when the effect is – by chance – relatively small, the results will be non-significant. Thus, given that an effect was significant, the measured effect size is probably larger than the actual effect size, and subsequent measurements will find smaller effects due to the familiar phenomenon of regression to the mean. The lower the statistical power, the more the effect size will be inflated.

## LOW-POWER, SMALL EFFECT SIZE

A number of findings suggest that the statistical power in psychology and neuroscience experiments is typically low. According to multiple meta-analyses, the statistical power of a typical psychology or neuroscience study to detect a medium-sized effect (defined variously as  $r = 0.3$ ,  $r = 0.4$ , or  $d = 0.5$ ) is approximately 0.5 or below (Cohen, 1962; Sedlmeier and Gigerenzer, 1989; Kosciolek and Szymanski, 1993; Bezeau and Graves, 2001). In applied psychology, power for medium effects is closer to 0.7, though it remains low for small effects (Chase and Chase, 1976; Mone et al., 1996; Shen et al., 2011). Nonetheless, many effects of interest in psychology are small and thus typical statistical power may be quite low. Field et al. (2009) report an average power of 0.2 in a meta-analysis of 68 studies of craving in addicts and attentional bias. In a heroic meta-analysis of 322 meta-analyses in social psychology, Richard et al. (2003) report that the average effect size was  $r = 0.21$ . To achieve power of 0.8 would require the average study to have 173 participants (in terms of medians:  $r = 0.18$ ,  $N = 237$ ), already far larger than typical sample size. Nearly 1/3 of the effect sizes reported were  $r = 0.1$  or less, requiring  $N = 772$  to achieve power of 0.8.

All else being equal, low statistical power would increase the proportion of significant results that are spurious. For instance, suppose researchers are investigating a hypothesis that is equally likely to be true or false (the prior likelihood of the null hypothesis is 50%), using methods with statistical power = 0.8. In this case, 6% of significant results will be false positives (True positives:  $0.5 \times 0.8 = 0.4$ ; False positives:  $0.5 \times 0.05 = 0.025$ ; Ratio:  $0.025/0.425 = 0.059$ ). If Power = 0.2, this increases to 20%. If the prior likelihood of the null hypothesis is 90% (i.e., if an effect would be surprising, or when data-mining), the false positive rate will be 69% (for additional discussion, see Yarkoni, 2009; for other problems associated with small power, see Tversky and Kahneman, 1971).

## FAILURE TO ACCOUNT FOR MULTIPLE COMPARISONS

If one tests for 10 different possible effects in each experiment, the chance of finding at least one significant at the  $p = 0.05$  level even when no effect actually exists is  $1 - 0.95^{10} = 0.4$ . Since experiments with large numbers of comparisons are often entirely exploratory, where there is no strong a priori reason to believe that any of the investigated effects exist, the false positive rate may approach 100% for data-mining studies with large datasets.

## DATA-PEEKING AND CONTINGENT STOPPING OF DATA COLLECTION

Many researchers compile and analyze data prior to testing a full complement of subjects. There is nothing wrong with this, so long as the decision to stop data collection is made independent of the results of these preliminary analyses, or so long as the final result is then replicated with the same number of subjects. Unfortunately, the temptation to stop running participants once significance is reached – or to run additional participants if it has not been reached – is difficult to resist. This data-peeking and contingent stopping has the potential to significantly increase the false positive rate (Feller, 1940; Armitage et al., 1969; Yarkoni and Braver, 2010). Even if the null hypothesis is true, a researcher who tests for significance after every participant has a 25% chance of finding a significant result with 20 or fewer participants (if the underlying distribution is normal; the analogous numbers are 19.5% for exponential distributions and 11% for binomial distributions; Armitage et al., 1969). This issue may be mitigated by use of alternative statistical tests, such as Bayesian statistics (Edwards et al., 1963), but such statistics have not been widely adopted.

## NEWSPWORTHINESS BIAS

Researchers are more likely to submit – and editors more likely to accept – “newsworthy” or surprising results. Spurious results are likely to be surprising, and thus are likely to be over-represented in published reports. Consistent with this claim, there is some evidence that highly cited papers are less likely to replicate (Ioannidis, 2005a) and that publication bias affects high-impact journals more severely (Ioannidis, 2005a; Munafo et al., 2009).

## HOW REPLICABLE ARE PUBLISHED STUDIES?

Several studies have found low rates of replicability across multiple scientific fields. Ioannidis (2005a) found that of 34 highly cited clinical research studies for which replication attempts had been published, seven (20%) did not replicate. Boffetta et al. (2008)



report a number of cases in which reports of significant cancer risk factors did not replicate. Recent studies have reported that relatively few genetic association links can be replicated (Ioannidis et al., 2001, 2003; Hirschhorn et al., 2002; Lohmueller et al., 2003; Trikalinos et al., 2004).

Likewise, several studies have found that initial reports of effect size are often exaggerated. This has been noted in medicine (Ioannidis et al., 2001, 2003; Trikalinos et al., 2004; Ioannidis, 2005a; but see Gehr et al., 2006), with similar declines in effect size reported in ecological and evolutionary biology (Jennions and Møller, 2002a,b). In the most extreme example, Dewald et al. (1986) reanalyzed the datasets underlying published studies in economics and were unable to fully replicate the analyses for seven of nine (78%).

Less is known about replication rates in psychology and neuroscience. In a series of five meta-analyses of fMRI studies, Wager and colleagues estimated that between 10 and 40% of activation peaks are false positives (Wager et al., 2007, 2009). While there seem to be few systematic surveys within psychology, some published effects are known not to replicate, such as the initial finding that violent video games increase violent behavior (Ferguson and Kilburn, 2010), various claims about the relationship between birth order and personality (Ernst and Angst, 1983; Harris, 1998; but see: Kristensen and Bjerkedal, 2007; Hartshorne et al., 2009), and a range of gene/environment interactions (Flint and Munafò, 2009).

In order to add to our knowledge of replicability rates in psychology and related disciplines, we surveyed 49 researchers in these disciplines, who reported a total of 257 attempted replications of published studies (for details, see Appendix). Only 127 (49%) fully replicated the original findings. This low rate was not driven by a small number of researchers attempting a large number of poor quality replications: both the mean and median replication success rates were 50%, with 77% of researchers reporting at least one attempted replication. Thus, the results of this survey suggest that replication rates within psychology and related disciplines are undesirably low, in accordance with the low rates of replicability found in many other fields.

## INCENTIVES IN PUBLICATION

As reviewed above, a number of factors promote low replicability rates across a range of fields. These problems are reasonably well known, and in many cases solutions have been proposed, such as use of different statistical methods and self-replication prior to publication. However, in spite of these solutions, evidence suggests that replicability remains low and thus that the proposed solutions have not been widely adopted. Why would this be the case? We propose that the incentive structure of the current system diminishes the ability and tendency of researchers to adopt these solutions. Namely, current methods of judging paper, researcher, and journal quality fail to take replicability into account, and in effect incentivize publishing spurious results.

## QUANTIFYING RESEARCH QUALITY

There are three primary *quantitative* criteria by which researchers are judged: their number of publications, the impact factor of the journals in which the publications appear, and the number of citations those papers receive. These quantitative values are a major

consideration in the awarding of grants, hiring, and tenure. Journals are similarly judged in terms of citation counts, which are compiled to calculate journal impact factors. Unfortunately, these metrics of quality tend to disincentivize taking additional steps to ensure the reliability of published findings, for several reasons.

Firstly, eliminating false positives means publishing fewer papers, since null results are difficult to publish. Second, ensuring that effect sizes are not inflated means reporting results with smaller effect sizes, which may be seen as less interesting or less believable. Third, as discussed above, spurious results are more likely to be surprising and newsworthy. Thus, eliminating spurious results disproportionately eliminates publications that would be widely cited and published in top journals.

These drawbacks are compounded by the fact that many of the improved practices that ensure replicability take time and resources. Learning to use new statistical methods often requires substantial effort. Increasing an experiment's statistical power may require testing more participants. Eliminating stopping of data collection contingent on significance level (data-peeking) also means erring on the side of testing more participants. Perhaps the best insurance against false positives is pre-publication replication by the authors. All these strategies take time.

In addition, there is relatively little cost associated with publishing unreliable results, as failures to replicate are rarely published and not systematically tracked. As a result, knowledge of the replicability of results mainly travels via word-of-mouth, through specific personal interactions at conferences and meetings. There are obvious concerns about the reliability of such a system, and there is little evidence that this system is particularly effective. We are aware of several cases in which a researcher invested months or years into unsuccessfully following up on a well-publicized effect from a neighboring subfield, only to later be told that it is "well-known" that the effect does not replicate.

Moreover, even when a failure-to-replicate is published, the results often go unnoticed. For example, a meta-analysis by Maraganore et al. (2004) concluded that UCHL1 is a risk-factor for Parkinson's Disease. Subsequent more highly powered meta-analyses overturned this result (Healy et al., 2006). Nonetheless, Maraganore et al. (2004) has been cited 70 times since 2007 (Google Scholar, May 10, 2011), much to the dismay of the senior author of the study (Ioannidis, 2011). Even papers retracted by the authors remain in circulation. In 2001, two papers were retracted by Karen Ruggiero (Ruggiero and Marx, 1999; Ruggiero et al., 2000). Nonetheless, 10 of the 22 citations to these papers were made in 2003 or later (Google Scholar, April 25, 2011). Similarly, though Lerner requested the retraction of Lerner and Gonzalez (2005) in 2008, the paper has been cited five times in 2010–2011 (Google Scholar, April 25, 2011).

It follows that researchers who take additional steps to ensure the quality of their data will ultimately spend more time and resources on each publication and, all else equal, will end up with fewer, less-often-cited papers in lower-quality journals. In the same way, journals that adopt more stringent publication standards may drive away submissions, particularly of the surprising, newsworthy findings that are likely to be widely cited. Certainly, the vast majority of researchers and editors are internally motivated to publish real, reliable results. However, we also cannot continue

practicing science without jobs, grants, and tenure. This situation sets up a classic Tragedy of the Commons (Hardin, 1968): While it is in everyone's collective interest to adopt strategies to improve replicability, the incentives for any *individual* researcher run the other direction.

### ESCAPING THE TRAGEDY OF THE COMMONS

Individuals can solve the Tragedy of the Commons by adopting common rules or changing incentive structures. To give a recent example, Jaeger (2008), Baayen et al. (2008), and others convinced many language processing researchers to switch from ANOVAs to mixed effects models, in part by convincing editors and reviewers to insist on it. In this case, collective action motivated widespread adoption of an improved method of analysis.

In a similar way, collective action is needed to solve the problem of low replicability: Because the incentive structure of the current system penalizes any member of the community who is an early adopter of reforms, an organized community change is needed. Instead of maintaining a system in which individual incentives (publish as often as possible) run counter to the goals of the group (maintain the integrity of the scientific literature), we can change the incentives by placing value on replicability directly. To do this, we propose tracking the replicability of published studies, and evaluating the quality of work post-publication partly on this basis. By tracking replicability, we hope to provide concrete incentives for improvements in research practice, thus allowing the widespread adoption of these improved practices.

### REPLICATION TRACKER: A PROPOSAL

Below, we lay out a proposal for how replications might be tracked via an online open-access system tentatively named *Replication Tracker*. The proposed system is not yet constructed; our aim in this proposal is to spur necessary discussion on the implementation of such a system. We first describe the core components of such a system. We then discuss in more depth issues that arise, such as motivating participation, aggregating information, and ensuring data quality.

### CORE ELEMENTS OF THE REPLICATION TRACKER

In a system such as Google Scholar, each paper's reference is presented alongside the number of times that paper has been cited, and each paper is linked to a list of the papers citing that target paper. Replication Tracker would function in a similar manner, except that it would be additionally indexed by specialized citations that link papers based on one attempting to replicate the other. Thus, each paper's reference would appear alongside not only a citation count, but an attempted replication count and information about the paper's replicability.

Replication Tracker's attempted replication citations are termed *Replication Links* (henceforth *RepLinks*). Each RepLink is tagged with metadata, answering the question: To what extent are these findings strong evidence that the target paper does or does not replicate? This metadata takes the form of two numerical ratings: a *Type of Finding Score*, running from +2 (fully replicated) to -2 (fully failed to replicate); and a *Strength of Evidence Score*, running from 1 (weak evidence) to 5 (strong evidence). These ratings, as well as the RepLinks themselves, could be produced through a

variety of methods; we suggest crowd-sourcing from the scientific community, as outlined below.

For replications to be tracked, they must be reported. As discussed above, many replication attempts remain unpublished. Thus, Replication Tracker would be paired with an online, open-access journal devoted to publishing Brief Reports of replication attempts. After a streamlined peer review process, these Brief Reports would be published and connected to the papers they replicate via RepLinks in the Replication Tracker.

This system will ultimately form a rich dataset, consisting of RepLinks between attempted replications and the original findings. Each RepLink's ratings would indicate the type and strength of evidence of the findings. These ratings would be aggregated, and used to compute statistics on replicability. For instance, the system could summarize the data for each paper in terms of a *Replicability Score* [e.g., 15 attempted replications, Replicability Score: +1.7 (Partial Replication), Strength of Evidence: 4 (Strong)], much as citation indices score papers based on citation counts (e.g., cited by 15). These numbers would allow researchers to both get an initial impression of a finding's replicability at a glance, and quickly click through to the original sources for further detail. In addition, Replicability Scores could be aggregated for each journal, which could be used alongside the existing Impact Factor to evaluate the quality of journals.

### STRUCTURE AND CONTENT OF RepLinks

RepLinks must, minimally, link a replication attempt with its target paper, note whether the finding was replication or non-replication, and note the strength of evidence for this finding.

There are many factors that enter into these decisions. For instance, a particular attempted replication may have investigated all of the findings in the target paper, or may have only attempted to replicate some subset. The findings may be more similar or less similar as well: All effects may have successfully replicated, or none; or some findings may have replicated while others did not. In addition, whether a replication serves as strong evidence of the replicability or non-replicability of the original finding depends on the extent of similarity of the methods used, and whether the attempt had high or low statistical power.

We propose capturing these issues in two ratings. The first rating, termed the *Type of Finding* rating, would take into account two factors: Whether all or only a subset of the target papers' findings were investigated; and whether all, none, or some of the attempted replications were successful. On this Type of Finding scale, -2 would denote a total non-replication (all findings investigated; none replicated); -1 a partial non-replication (some subset of findings investigated; none of those investigated replicated); 0 would denote mixed results (of the findings investigated, some replicated, and others did not); 1 a partial replication (some subset of findings investigated; all of those investigated replicated); and 2 a total replication (all findings investigated; all replicated).

The second rating would be a *Strength of Evidence* rating, scored on a 1–5 scale. This rating would take into account the remaining two factors: the extent to which the methods are similar between the target paper and the RepLinked paper, and the power of the replication attempt. Thus a score of 5 reflects a high-powered attempt with as-close-as-possible methods, while 1

reflects a low-powered attempt with relatively dissimilar methods. When a replication attempt is extremely low-power or uses substantially different methods, it would not be assigned a RepLink at all.

### WHO CREATES AND RATES REPLINKS?

The ratings described above involve a number of difficult determinations. Given that no two studies can have exactly identical methods, how similar is similar enough? How does one determine whether a study has sufficient statistical power, given that the effect's size is itself under investigation?

To make these determinations, we turn to those individuals most qualified to make them: researchers in the field. Crowd-sourcing has proven a highly effective mechanism of making empirical determinations in a variety of domains (Giles, 2005; Law et al., 2007; von Ahn and Dabbish, 2008; von Ahn et al., 2008; Bederson et al., 2010; Yan et al., 2010; Doan et al., 2011; Franklin et al., 2011). Researchers would form the user base of the system, and any user could submit a RepLink, as well as a Type of Finding and Strength of Evidence score for a RepLink. When submitting these materials, users could also optionally comment on each RepLink, providing a more detailed description of how the methods or results of the RepLinked paper differed from the target paper, or offering interpretations of discrepancies. These comments would be optionally displayed alongside each users' individual ratings, for readers looking for additional detail (Figure 4).

The system also utilizes multiple moderators. These moderators would take joint responsibility for tending the RepLinks and Brief Reports (see below) on papers in their subfields. Moderators would be scientists, and could be invited (e.g., by the founding members), although anyone with publications in the field could apply to be a moderator.

In submitting and rating RepLinks, researchers may disagree with one another as to the correct Type of Finding or Strength of Evidence ratings for a given RepLink, or may disagree as to whether two papers are sufficiently similar as to qualify as a replication attempt. Users who agree with an existing rating may easily second it with a thumbs-up, while users who disagree with the existing ratings may submit their own additional ratings. Users who believe that the papers in question do not qualify as replications may flag the RepLink as irrelevant (RepLinks that have been flagged a sufficient number of times would no longer be used to calculate Replicability Scores, though these suppressed RepLinks would be visible under certain search options). These ratings would be combined together using crowd-sourcing techniques to determine the aggregate Type of Finding and Strength of Evidence scores for a given RepLink (see below).

### AGGREGATION, AUTHORITY, AND MACHINE LEARNING

Data must be aggregated by this system at multiple levels. First, multiple ratings for a given RepLink must be combined into aggregate Type of Finding and Strength of Evidence ratings for that RepLink. Second, where a single target paper has been the subject of multiple replication attempts, the different RepLinks must be aggregated into a single Replicability Score and Strength Score for that target paper. In the same way, scores may be combined

across multiple papers to determine aggregate replicability across a literature, an individual researcher's publications, or a journal.

Aggregates need not be mere averages. How to best aggregate ratings across multiple raters is an active area of research in machine learning (Albert and Dodd, 2004; Adamic et al., 2008; Snow et al., 2008; Callison-Burch, 2009; Welinder et al., 2010). Type of Finding ratings for an individual RepLink may be weighted by their associated Strength of Evidence scores, as well as how many thumbs-up they have received.

In addition, ratings from certain users would be weighted more heavily than others, as is done in many rating aggregation algorithms (e.g., Snow et al., 2008). There are many mechanisms for doing so, such as downgrading the authority of users whose RepLinks are frequently flagged as irrelevant, and assigning greater authority to moderators. The best system of weighting and aggregating RepLinks is an interesting empirical question. We see no reason it must be set in stone from the outset; the best algorithms may be determined through new research in machine learning. To that end, the raw rating dataset would be made available to those working in machine learning and related fields.

### A NOTE ON CONVERGING RESULTS

Only strict replications, not convergent data from different methods, will be tracked in the proposed system. This may seem counter-intuitive, since tracking converging results is crucial for determining which theories are most predictive. However, the goal of the proposed system is not to directly evaluate which *theories* are right, but to determine which *results* are right – that is, which patterns of data are reliable. Consider that while converging results may suggest that the original finding replicates, *diverging* results may only indicate that the differences in the methodologies were meaningful. For this reason, we focus solely on tracking strict replications. We believe that evaluating the complex theoretical implications of a large body of data is best handled by researchers themselves (i.e., when writing review papers), and is likely not feasible with an automated system.

### AUTHENTICATION AND LABELING OF AUTHORS' RATINGS AND COMMENTS

Registering for the system and submitting RepLinks would not require authenticating one's identity. However, authors of papers could choose to have their identities authenticated in order to have comments on their own papers be marked as author commentaries (many RepLinks will almost certainly be submitted by authors, as they are most invested in the issues involved in replication of their own studies).

Identity authentication could be accomplished in multiple ways. For instance, a moderator could use the departmental website to verify the author's email address and send a unique link to that email address. Clicking on that link would enable the user to set up an authenticated account under the users' own name. Moderator's identities could be authenticated in a similar manner.

### SELECTION OF MODERATORS

Although any user can contribute to Replication Tracker, moderators play several additional key roles. First, they evaluate submitted Brief Reports, and submit the initial RepLinks for any accepted

Brief Report. Similarly, when new RepLinks are submitted, moderators are notified and can flag irrelevant RepLinks or submit their own ratings. Thus, it is important that (a) there are enough moderators, and (b) the moderators are sufficiently qualified. In the case of moderator error, the Replication Tracker contains numerous ways by which other moderators and users can override the erroneous submission (submitting additional RepLink scores; flagging the erroneous RepLink, etc.). In order to recruit a sufficient number of moderators, we suggest allowing existing moderators to invite additional moderators as well as allowing researchers to apply to be moderators. Moderators could be selected based on objective considerations (number of publications, years of service, etc.), subjective considerations (by a vote of existing moderators), or both.

### RETRACTIONS

The Replication Tracker system is also ideally suited to tracking retractions. Retractions may be submitted by users as a specially marked type of RepLink, which would require moderator approval before posting. Retracted studies would appear with the tag *RETRACTED* in any search results, and automatically be excluded from calculations of Replicability Scores. As a safeguard against incorrect flags, any time a study is flagged as retracted, all other moderators would be notified, and the flag could be revoked if found inaccurate.

### BRIEF REPORTS

The efficacy of Replication Tracker is limited by the number of published replication attempts. As discussed above, both successful replications and null results are difficult to publish, and often remain undocumented. Thus, we propose launching an open-access journal that publishes all and any replication attempts of suitable quality.

Unlike full papers elsewhere, these *Brief Reports* would consist of the method and results section only. This greatly reduces the cost of either writing or reviewing the report. The Brief Report must also be submitted with one or more RepLinks, specifying what exactly is being replicated. Particularly for non-replications, authors of Brief Reports can use the comments on the RepLinks to discuss why they think the replication failed (low-power in the original study, etc.).

Review of Brief Reports would be handled by moderators. When a Brief Report is submitted, all moderators of that sub-field would be automatically emailed with a request to review the proposed post. The review could then be “claimed” by any moderator. If no one claims the post for review within a week, the system would then automatically choose one of the relevant moderators, and ask if they would accept the request to review; if they decline, further requests would be made until someone agreed to review. Authors would not be able to be the sole moderator/reviewer for replications of their own work. As in the PLoS model, the moderator could evaluate the *Brief Report* alone or solicit outside review(s).

The presumption of the review process would be acceptance. Brief Reports would be returned for revision when appropriate, as in the case of using inappropriate statistical tests; but would only be rejected if the paper does not actually qualify as a replication

attempt (based on the criteria discussed above). In the latter case, authors of Brief Reports could appeal the decision, which would then be reviewed by two other moderators. On acceptance, the Brief Report would be published online in static form with a DOI, much like any other publication, and thus be part of the citable, peer reviewed record. The appropriate RepLinks would be likewise added to Replication Tracker. As with any RepLink, these could be suppressed if flagged as irrelevant a sufficient number of times (see above). Thus, while publication in Brief Reports is permanent (barring retractions), incorporation into Replication Tracker is always potentially in flux – as is appropriate for a post-review evaluation process.

### THE EXPERIENCE OF USING REPLICATION TRACKER: A STEP-BY-STEP GUIDE

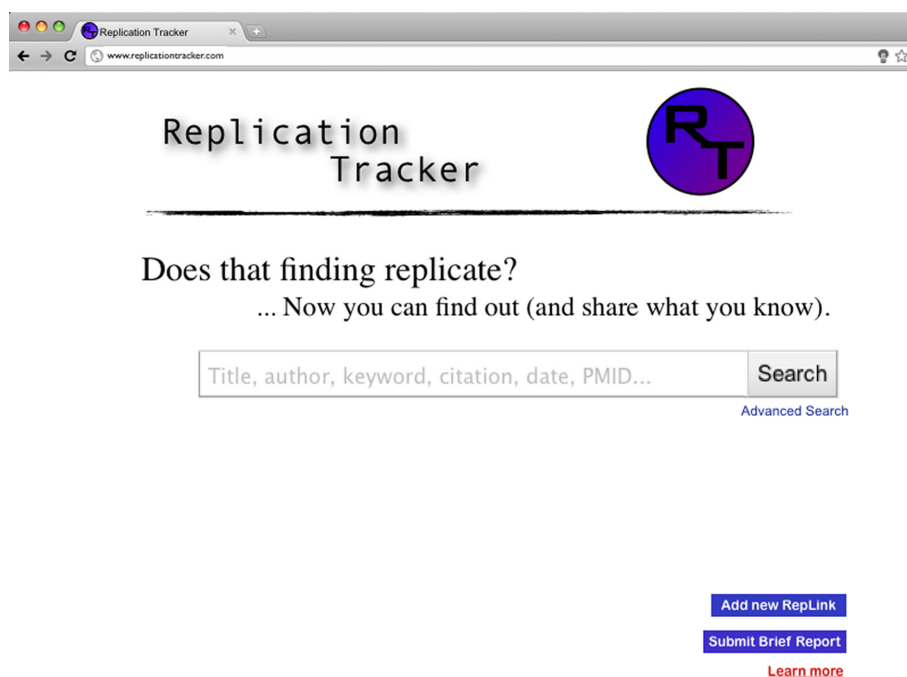
As in any literature database, users would begin by using a search function (either simple or advanced) to locate a paper of interest (**Figure 1**). This search would bring up a list of references, in a format similar to Google Scholar. However, in addition to the citation count provided by Google Scholar, the system would provide three additional values: The number of replication attempts documented, the paper’s Replicability Score, and the Strength of Evidence score (**Figure 2**). As described above, the Replicability Score would hold a value from  $-2$  to  $+2$ , with negative values denoting evidence of non-replication, zero denoting mixed findings, and positive values evidence of successful replication.

The user would then click on a reference from the list to bring up more detailed information about that target paper (**Figure 3**). The target paper’s reference would appear at the top of the page, along with the number of attempted replications documented, Replicability Score for that paper, and the Strength of Evidence score. Below these aggregate measures would be a list of the RepLinks, represented by a citation of the RepLinked paper, the aggregate Type of Finding score and Strength of Evidence score for that RepLink, and the number of users who have rated that RepLink. An additional button would allow users to add their own ratings or flag the RepLink as irrelevant.

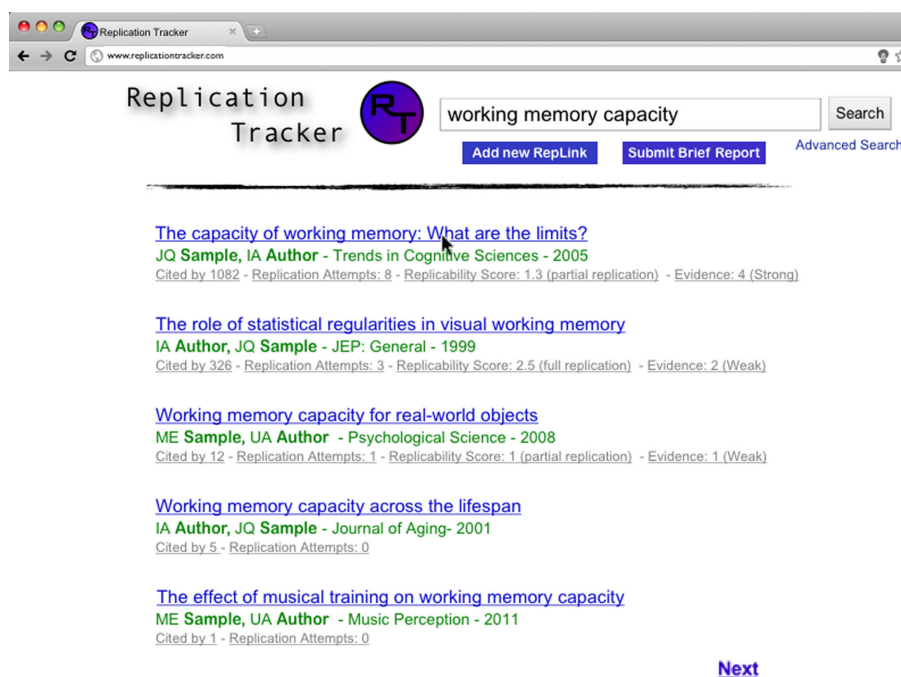
Information about each RepLink could be expanded, to show each individual rating along with that users’ associated comments, if any (**Figure 4**). Users could agree with an existing rating via a thumbs-up button. Ratings and comments would be labeled with the username of the poster; for authenticated accounts, they could optionally be labeled with the individuals’ real name. Comments by authors who have chosen to authenticate their account under their real names would be labeled as such.

### ISSUES FOR FURTHER DISCUSSION

The Replication Tracker would serve several functions. First, it would enable a new way of navigating the literature. Second, we believe it would motivate researchers to conduct and report attempted replications, helping correct biases in the literature such as the file-drawer problem. Third, it will vastly improve access to and communication regarding replication attempts. Perhaps most importantly, it would help incentivize and reward costly efforts to ensure replicability pre-publication, helping to mitigate a Tragedy of the Commons in scientific publishing.



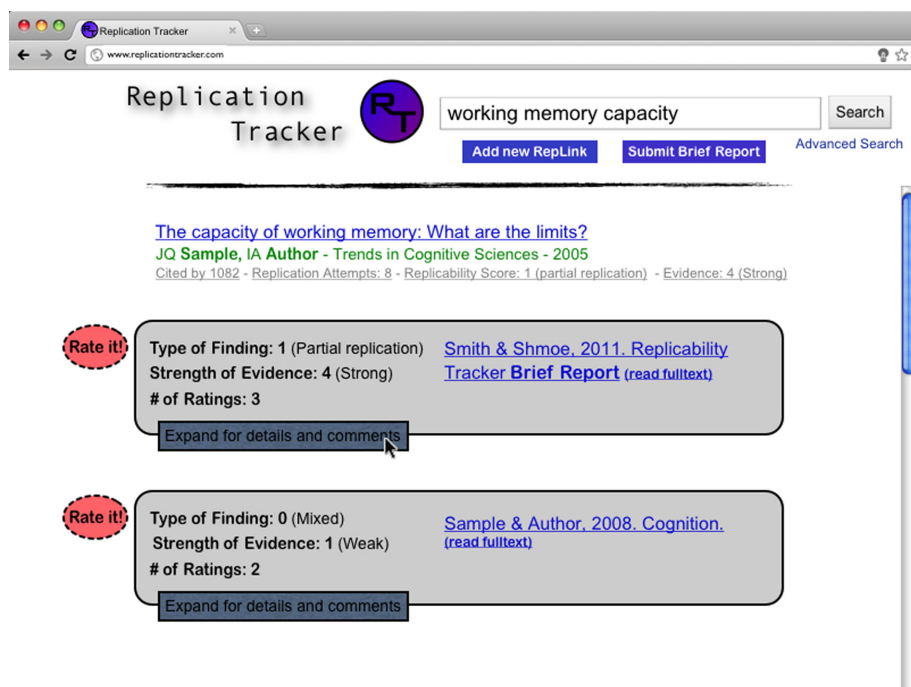
**FIGURE 1 | Replication tracker: search window.** Much like any other paper index, Replication Tracker would allow the user to search for papers by author, keyword, and other typical search terms.



**FIGURE 2 | Replication tracker: example search results.** Results of a search query list relevant papers, along with number of citations and information about the paper's replicability. This information consists of the number of attempted replications reported to the system, a summary statistic of

whether the finding successfully replicates or fails to replicate ("Replicability Score"), and a summary statistic of the strength of the evidence. These numbers are derived from RepLinks, data which is crowd-sourced from users and moderators (**Figure 3**).





**FIGURE 3 | Replication tracker: search results expansion, showing RepLinks for a target paper.** Each RepLink represents an attempted replication. Again, the degree of success of the replication

("replication type") and strength of the evidence is noted. These are determined by aggregating determinations made by individual users (Figure 4).



**FIGURE 4 | Replication tracker: expansion of a RepLink, showing ratings by individual readers, which are summarized in Figure 3.** Users are also able to add comments, explaining their determinations, or flag posts as irrelevant, prompting review by moderators.

However, in addition to these potential benefits, tracking, and publishing replication attempts raises non-trivial issues, and has the potential for unintended consequences. We consider several such concerns below and discuss how these concerns may be addressed or allayed.

### GETTING THE SYSTEM OFF THE GROUND

The usefulness of the database for tracking replicability will be a function of the amount of replication information added to it in the form of RepLinks, metadata information, and Brief Reports. This will require considerable participation by a broad swath of the research community. Because researchers are more likely to contribute to a system that they already find useful, an important determiner of success will be the ability to achieve a critical mass of such information. We have considered several ways of increasing the likelihood that the system quickly reaches critical mass.

First, there should be a considerable number of founding members, so that a wide range of researchers are engaged in the project prior to launch. This will not only help with division of labor, but will also help clarify the many design decisions that go into creating the details of the system. The more diverse the founding group is, the more likely the final system will be acceptable to researchers in multiple fields and disciplines. This paper serves as a first step in starting the needed dialog.

Second, we suggest concentrating on first reaching critical mass for a few select subfields of psychology and neuroscience, instead of simultaneously attempting to obtain critical mass in all fields of science at once. In order to reach critical mass within the first few subfields, we suggest that prior to the public launch of Replication Tracker, founding members conduct targeted replicability reviews of specific literatures within those subfields, writing RepLinks and soliciting Brief Reports during the process. These data would be used to write review papers, which would be published in traditional journals. These review papers would be useful publications in and of themselves and would help demonstrate the empirical value of tracking replications. This would help recruit additional founders, moderators and funding – all while major components are added to the database. Only once enough coverage of the literatures within those subfields has been achieved would Replication Tracker be publically launched.

In addition to tracking published replications, the proposed system attempts to ameliorate the file-drawer problem by allowing researchers to submit Brief Reports of attempted replications. Several previous attempts have been made to publish null results and replication attempts (e.g., Journal of Articles in Support of the Null Hypothesis; Journal of Negative Results in Biomedicine) often with low rates of participation (JASNH has published 32 papers since its launch in 2002). Nonetheless, we believe several aspects of our system would motivate increased participation. Firstly, the format of Brief Reports significantly decreases the time commitment of preparation, as the Reports consist of the method and results section only. Second, these Brief Reports will not only be citable, but will also be highly findable, as they will be RepLinked to the relevant published papers. Thus we expect these Reports to have some value, perhaps equivalent to a conference paper or poster. We believe that the combination of lesser time investment and increased value will lead to increased rates of submission.

### WHAT IS THE RIGHT UNIT OF ANALYSIS?

Because each paper may include multiple findings that differ in replicability, there is a good argument to be made that what should be tracked is the replicability of a given result. We propose tracking the replicability of papers instead, for several reasons.

The first reason is one of feasibility. We believe that tracking each finding separately would be infeasible, as what counts as an individual finding may be subjective, and the vast number of units of analysis even within a single paper becomes prohibitive. An intermediate level would be to track individual experiments. However, publication formats do not always include separate headings for each individual experiment (e.g., *Nature*, *Current Biology*), and even a single experiment may include multiple components with differences in replicability.

Secondly, even organizing the system at the level of experiment will not allow an aggregated replicability score to capture every nuance of the scientific literature. It will always be necessary for the reader to examine written information for more detail, including the full text of the RepLinked papers. For these detail-oriented readers, the proposed system provides a novel way to navigate through published work (by following RepLinks to find and read papers with attempted replications) and an efficient way to view comments on each of these papers (Figure 4). Such a system is most intuitive and navigable when organized at the level of the paper itself.

### ARE SUFFICIENT NUMBERS OF REPLICATIONS CONDUCTED?

The rate of published replications appears to be low: For instance, over a 20-year period, only 5.3% of 701 publications in nine management journals included attempts to replicate previous findings (Hubbard et al., 1998). While we believe Replication Tracker would lead to increased numbers of published replications, we must consider whether Replication Tracker would be useful if the number of published replications remains low. Certainly, many papers will simply never be replicated, and many others will only have one reported replication attempt.

We do not believe these issues undermine the utility of Replication Tracker for several reasons. First, the findings which are of broadest interest to the community are likely the very same findings for which the most replications are attempted. Thus, while many low-impact papers may lack replication data, the system will be most useful for the papers where it is most needed. Secondly, even low numbers of replications are often sufficient: because spurious results are unlikely to replicate, even only a handful of successful replications significantly increases the likelihood that a given finding is real (Moonesinghe et al., 2007). Finally, we note that even sparse replicability data is useful when aggregating over large numbers of papers, for instance, when producing aggregate Replicability Scores for journals. Similarly, it would be possible to aggregate across studies within individual literatures or using particular methods. For these aggregate scores, sparse data does not present a problem.

### WOULD TRACKING REPLICABILITY STIFLE NOVEL SCIENTIFIC FIELDS?

Commenters on the present paper have suggested that since new fields may still be designing the details of their methods, and may be less sure of what aspects of the method are necessary to correctly

measure the effects under investigation, their initial results may appear less replicable. In this case, using replicability scores as a measure of paper, researcher, and journal quality – one of our explicit aims – could potentially stifle new fields of enquiry.

This is an important concern if true. We do not know of any systematic empirical data that would adjudicate the issue. However, we suspect that other factors may systematically increase replicability in new lines of inquiry. For example, young fields may focus on larger effects, with established fields focusing on increasingly subtle effects over time (cf Taubes and Mann, 1995). Additionally, in the case that subtle methodological differences prevent replication of results, Replication Tracker may actually aid researchers in identifying the relevant issues more quickly, spurring growth of the novel field.

We additionally note that it is not our intention that replicability become the sole criteria by which research quality is measured, nor do we think that is likely to happen. New fields are likely to generate excitement and citations, which will produce their own momentum. The goal is that replicability rates be considered in addition.

#### **WOULD REPLICATION TRACKER UNDERESTIMATE REPLICABILITY?**

Commenters on the present paper have also suggested several ways in which Replication Tracker might underestimate replicability. Underestimating the replicability of a field could undermine both scientists' and the public's confidence in the field, leading to decreased interest and funding.

##### ***Null effect bias***

Researchers may be more motivated to submit non-replications to the system as Brief Reports, while successful replications would languish in file-drawers. We suspect that this problem would disappear as the system gains popularity: Researchers typically attempt replications of effects that are crucial to their own line of work and will find it useful to report those replications in order to have their own work embedded in a well-supported framework. Moreover, many replication attempts are conducted by the authors of the original study, who will be intrinsically motivated to report successful replications in support of their own work. Nonetheless, this is an issue that should be evaluated and monitored as Replication Tracker is introduced, so that adjustments can be made as necessary.

##### ***Unskilled replicators***

Another concern is that if on average the researchers that tend to conduct large numbers of strict replications are less skilled than the original researchers, this could lead to non-replications due to unknown errors. If this is the case, this issue could be compensated for in two ways. First, as Replication Tracker and Brief Reports raise the profile of replication, more skilled researchers may begin to conduct and report more replications. Second, as discussed above, there are numerous machine learning techniques to identify the most reliable sources of information. These techniques could be applied to mitigate this issue, by discounting replication data from users that have not been reliable sources of information in the past.

##### ***Spurious non-replications***

Since the statistical power to detect an effect is never 1.0, even true effects sometimes do not replicate. High-profile papers in

particular will be much more likely to be subject to replication attempts; since some replications even of real effects will fail, high-profile papers may be unfairly denigrated. This issue is compounded if typical statistical power in that literature is low, making replication improbable.

These issues can be dealt with directly in Replication Tracker, by appropriately weighing this probabilistic information. Recall that Replication Tracker provides both a Replicability Score, indicating whether existing evidence suggests that the target paper replicates, as well as a Strength of Evidence Score. A single non-replication – particularly one with only mid-sized power – is not strong evidence for non-replicability, and this should be reflected in the Strength Score. Replication attempts with low-power should not be RepLinked at all. If 8 of 10 replication attempts succeed – consistent with statistical power of 0.8 – that should be counted as strong evidence of replicability.

#### **WILL TYPE II ERROR INCREASE?**

Finally, we must consider whether the changes people will make to their work will actually lead to an increased  $d'$  (ability to detect true effects) or whether these changes will simply result in a trade-off: researchers may eliminate some false positives (Type I error) only at the expense of increasing the false negative rate (Type II error). It is an open question whether fields like psychology and neuroscience are currently at an optimal balance between Type I and Type II error, and Replication Tracker would help provide data to adjudicate this issue. Moreover, some of the potential reforms would almost certainly increase  $d'$ , like conducting studies with greater statistical power.

#### **LIMITATIONS TO EVALUATION BY TRACKING REPLICATIONS**

Replicability is a crucial measure of research quality; however, certain types of errors cannot be detected in by such a system. For instance, data may be misinterpreted, or a flawed method of analysis may be repeatedly used. Thus, while tracking replicability is an important component of post-publication assessment, it is not the only one needed. We have suggested presenting replicability metrics side-by-side with citation counts (Figure 2). Similarly, other post-publication evaluations, such as those described within other papers in this Special Topic, could be presented alongside these quantitative metrics.

While it is tempting to try to build a single system to track multiple aspects of research quality, we believe that constructing such a system will be extremely difficult, as different data structures are required to track each aspect of research quality. The Replication Tracker system, as currently envisioned, is optimized for tracking replications: The basic data structure is the RepLink, a connection between a published paper and a replication attempt of its findings. In contrast, to determine the truth value of a particular idea or theory, papers should be rated on how well the results justify the conclusions and linked to one another on the basis of theoretical similarity, not just strict methodological similarity. As such, we think that such information is likely best tracked by an independent system, which can be optimized accordingly. Ultimately, results from these multiple systems may then be aggregated and presented together on a single webpage for ease of navigation.

## CONCLUSION

In conclusion, we propose tracking replication attempts as a key method of identifying high-quality research post-publication. We argue that tracking and incentivizing replicability directly would allow researchers to escape the current Tragedy of the Commons in scientific publishing, thus helping to speed the adoption of reforms. In addition, by tracking replicability, we will be able to determine whether any adopted reforms have successfully increased replicability.

No measure of research quality can be perfect; instead, we aim to create a measure that is robust enough to be useful. Citation counts have proven very useful in spite of the metrics' many flaws as measures of a paper's quality (for instance, papers which are widely criticized in subsequent literature will be highly cited). We do not propose replacing citation counts with replicability measures, but rather augmenting the one with the other. Tracking replicability

## REFERENCES

- Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. (2008). "Knowledge sharing and yahoo answers: everyone knows something," in *Proceedings of the 17th International Conference on World Wide Web*, Beijing.
- Albert, P. S., and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 60, 427–435.
- Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J. R. Stat. Soc. Ser. A Stat. Soc.* 132, 235–244.
- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412.
- Bederson, B. B., Hu, C., and Resnik, P. (2010). "Translation by interactive collaboration between monolingual users," in *Proceedings of Graphics Interface*, Ottawa, 39–46.
- Bezeau, S., and Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *J. Clin. Exp. Neuropsychol.* 23, 399–406.
- Boffetta, P., McLaughlin, J. K., Vecchia, C. L., Tarone, R. E., Lipworth, L., and Blot, W. J. (2008). False-positive results in cancer epidemiology: a plea for epistemological modesty. *J. Natl. Cancer Inst.* 100, 988–995.
- Callahan, M. L., Wears, R. L., Weber, E. J., Barton, C., and Young, G. (1998). Positive-outcome bias and other limitations in the outcome of research abstracts submitted to a scientific meeting. *JAMA* 280, 254–257.
- Callison-Burch, C. (2009). "Fast, cheap, and creative: evaluating translation quality using Amazon's mechanical
- turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 286–295.
- Chase, L. J., and Chase, R. B. (1976). A statistical power analysis of applied psychological research. *J. Appl. Psychol.* 61, 234–237.
- Cohen, J. (1962). The statistical power of abnormal-social psychological research. *J. Abnorm. Soc. Psychol.* 65, 145–153.
- Cole, S. Jr., Cole, J. R., and Simon, G. A. (1981). Chance and consensus in peer review. *Science* 214, 881–886.
- Coursol, A., and Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: a note on meta-analysis bias. *Prof. Psychol. Res. Pr.* 17, 136–137.
- Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., and Andersson, G. (2010). Efficacy of cognitive-behavioral therapy and other psychological treatments for adult depression: meta-analytic study of publication bias. *Br. J. Psychiatry* 196, 173–178.
- Dewald, W. G., Thursby, J. G., and Anderson, R. G. (1986). Replication in empirical economics: the journal of money, credit and banking project. *Am. Econ. Rev.* 76, 587–603.
- Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., and Smith, H. (1987). Publication bias and clinical trials. *Control. Clin. Trials* 8, 343–353.
- Dickersin, K., Min, Y.-I., and Meinert, C. L. (1992). Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. *J. Am. Med. Assoc.* 267, 374–378.
- Doan, A., Ramakrishnan, R., and Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Commun. ACM* 54, 86–96.
- Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J. P., Simes, J., and Williamson, P. R. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE* 3, e3081. doi:10.1371/journal.pone.0003081
- Easterbrook, P. J., Berlin, J. A., Gopalan, R., and Matthews, D. R. (1991). Publication bias in clinical research. *Lancet* 337, 867–872.
- Edwards, W., Lindman, H., and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychol. Rev.* 70, 193–242.
- Ernst, C., and Angst, J. (1983). *Birth Order: Its Influence on Personality*. New York: Springer-Verlag.
- Eysenck, H. J., and Eysenck, S. B. (1992). Peer review: advice to referees and contributors. *Pers. Individ. Dif.* 13, 393–399.
- Feller, W. (1940). Statistical aspects of ESP. *J. Parapsychol.* 4, 271–298.
- Ferguson, C. J., and Kilburn, J. (2010). Much ado about nothing: the misestimation and overinterpretation of violent video game effects in eastern and western nations: comment on Anderson et al. (2010). *Psychol. Bull.* 136, 174–178.
- Field, M., Munafo, M. R., and Franken, I. H. A. (2009). A meta-analytic investigation of the relationship between attentional bias and subjective craving in substance abuse. *Psychol. Bull.* 135, 589–607.
- Flint, J., and Munafo, M. R. (2009). Replication and heterogeneity in gene x environment interaction studies. *Int. J. Neuropsychopharmacol.* 12, 727–729.
- Franklin, M., Kossman, D., Kraska, T., Ramesh, S., and Xin, R. (2011). CrowdDB: answering queries with crowdsourcing. *Paper Presented at the SIGMOD 2011*, Athens.
- Gehr, B. T., Weiss, C., and Porzolt, F. (2006). The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Med. Res. Methodol.* 6, 25. doi:10.1186/1471-2288-6-25
- Gerberg, A. S., Green, D. P., and Nickerson, D. (2001). Testing for publication bias in political science. *Polit. Anal.* 9, 385–392.
- Giles, J. (2005). Internet encyclopedias go head to head. *Nature* 438, 900–901.
- Hardin, G. (1968). The tragedy of the commons. *Science* 162, 1243–1248.
- Harris, J. R. (1998). *The Nurture Assumption: Why Children Turn out the Way That They Do*. New York: Free Press.
- Hartshorne, J. K., Salem-Hartshorne, N., and Hartshorne, T. S. (2009). Birth order effects in the formation of long-term relationships. *J. Individ. Psychol.* 65, 156–176.
- Healy, D. G., Abou-Sleiman, P. M., Casas, J. P., Ahmadi, K. R., Lynch, T., Gandhi, S., Muqit, M. M., Foltynie, T., Barker, T., Bhatia, K. P., Quinn, N. P., Lees, A. J., Gibson, J. M., Holton, J. L., Revesz, T., Goldstein, D. B., and Wood, N. W. (2006). UCHL1 is not a Parkinson's disease susceptibility gene. *Ann. Neurol.* 59, 627–633.
- Hirschhorn, J. N., Lohmueller, K. E., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genet. Med.* 4, 45–61.
- Hubbard, R., Vetter, D. E., and Little, E. L. (1998). Replication in strategic management: scientific testing for validity, generalizability, and usefulness. *Strateg. Manage. J.* 19, 243–254.

- Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *J. Am. Med. Assoc.* 294, 218–228.
- Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLoS Med.* 2, e124. doi:10.1371/journal.pmed.0020124
- Ioannidis, J. P. A. (2011). Meta-research: the art of getting it wrong. *Res. Syn. Methods* 1, 169–184.
- Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nat. Genet.* 29, 306–309.
- Ioannidis, J. P. A., Trikalinos, T. A., Ntzani, E. E., and Contopoulos-Ioannidis, D. G. (2003). Genetic associations in large versus small studies: an empirical assessment. *Lancet* 361, 567–571.
- Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446.
- Jennions, M. D., and Möller, A. P. (2002a). Publication bias in ecology and evolution: an empirical assessment using the “trim and fill” method. *Biol. Rev. Camb. Philos. Soc.* 77, 211–222.
- Jennions, M. D., and Möller, A. P. (2002b). Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proc. Biol. Sci.* 269, 43–48.
- Kosciulek, J. F., and Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counseling research. *Rehabil. Couns. Bull.* 36, 212–219.
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540.
- Kristensen, P., and Bjerkedal, T. (2007). Explaining the relation between birth order and intelligence. *Science* 316, 1717.
- Law, E., von Ahn, L., Dannenberg, R., and Crawford, M. (2007). TagATune: a game for sound and music annotation. *Paper Presented at the ISMIR*, Vienna.
- Lerner, J. S., and Gonzalez, R. M. (2005). Forecasting one's future based on fleeting subjective experiences. *Pers. Soc. Psychol. B.* 31, 454–466.
- Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., and Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* 33, 177–182.
- Mahoney, M. J. (1977). Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cognit. Ther. Res.* 1, 161–175.
- Maraganore, D. M., Lesnick, T. G., Elbaz, A., Chartier-Harlin, M.-C., Gasser, T., Kruger, R., Hattori, N., Mellick, G. K., Quattrone, A., Satoh, J.-I., Toda, T., Wang, J., Ioannidis, J. P. A., de Andrade, M., Rocca, W. A., and the UCHL1 Global Genetics Consortium. (2004). UCHL1 is a Parkinson's disease susceptibility gene. *Ann. Neurol.* 55, 512–521.
- Misakian, A. L., and Bero, L. A. (1998). On passive smoking: comparison of published and unpublished studies. *J. Am. Med. Assoc.* 280, 250–253.
- Mone, M. A., Mueller, G. C., and Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Pers. Psychol.* 49, 103–120.
- Moonesinghe, R., Khoury, M. J., and Janssens, C. J. W. (2007). Most published research findings are false—but a little replication goes a long way. *PLoS Med.* 4, e28. doi:10.1371/journal.pmed.0040028
- Munafo, M. R., Stothart, G., and Flint, J. (2009). Bias in genetic association studies and impact factor. *Mol. Psychiatry* 14, 119–120.
- Newton, D. P. (2010). Quality and peer review of research: an adjudicating role for editors. *Account. Res.* 17, 130–145.
- Olson, C. M., Rennie, D., Cook, D., Dickens, K., Flanagan, A., Hogan, J. W., Zhu, Q., Reiling, J., and Pace, B. (2002). Publication bias in editorial decision making. *JAMA* 287, 2825–2828.
- Peters, D. P., and Ceci, S. J. (1982). Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behav. Brain Sci.* 5, 187–195.
- Richard, F. D., Bond, C. F. Jr., and Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Rev. Gen. Psychol.* 7, 331–363.
- Ruggiero, K. M., and Marx, D. M. (1999). Less pain and more to gain: why high-status group members blame their failure on discrimination. *J. Pers. Soc. Psychol.* 77, 774–784.
- Ruggiero, K. M., Steele, J., Hwang, A., and Marx, D. M. (2000). Why did I get a 'D'? The effects of social comparisons on women's attributions to discrimination. *Pers. Soc. Psychol. B.* 26, 1271–1283.
- Saxe, R., Brett, M., and Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers. *Neuroimage* 30, 1088–1096.
- Sedlmeier, P., and Gigerenzer, G. (1989). Do studies of statistical power have an effects on the power of studies? *Psychol. Bull.* 105, 309–316.
- Sena, E. S., Worp, H. B. V. D., Bath, P. M. W., Howells, D. W., and Macleod, M. R. (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol.* 8, e1000344. doi:10.1371/journal.pbio.1000344
- Shen, W., Kiger, T. B., Davies, S. E., Rasch, R. L., Simon, K. M., and Ones, D. S. (2011). Samples in applied psychology: over a decade of research in a review. *J. Appl. Psychol.* 96, 1055–1064.
- Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). “Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks,” in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* Edinburgh, 254–263.
- Taubes, G., and Mann, C. C. (1995). Epidemiology faces its limits. *Science* 269, 164–169.
- Trikalinos, T. A., Ntzani, E. E., Contopoulos-Ioannidis, D. G., and Ioannidis, J. P. A. (2004). Establishment of genetic associations for complex diseases is independent of early study findings. *Eur. J. Hum. Genet.* 12, 762–769.
- Tversky, A., and Kahneman, D. (1971). Belief in the law of small numbers. *Psychol. Bull.* 76, 105–110.
- von Ahn, L., and Dabbish, L. (2008). General techniques for designing games with a purpose. *Commun. ACM* 51, 58–67.
- von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). reCAPTCHA: human-based character recognition via web security measures. *Science* 321, 1465–1468.
- Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4, 274–290.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., and van der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: Comment on Bem, 2011. *J. Pers. Soc. Psychol.* 100, 426–432.
- Wager, T. D., Lindquist, M., and Kaplan, L. (2007). Meta-analysis of functional neuroimaging data: current and future directions. *Soc. Cogn. Affect. Neurosci.* 2, 150–158.
- Wager, T. D., Lindquist, M. A., Nichols, T. E., Kober, H., and van Snellenberg, J. X. (2009). Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *Neuroimage* 45, S210–S221.
- Welinder, P., Branson, S., Belongie, S., and Perona, P. (2010). The multi-dimensional wisdom of the crowds. *Paper Presented at the Advances in Neural Information Processing Systems 2010*, Vancouver.
- Yan, T., Kumar, V., and Ganesan, D. (2010). “Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones,” in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, San Francisco.
- Yarkoni, T. (2009). Big correlations in little studies: inflated fMRI correlations reflect low statistical power—commentary on Vul et al. (2009). *Perspect. Psychol. Sci.* 4, 294–298.
- Yarkoni, T., and Braver, T. S. (2010). “Cognitive neuroscience approaches to individual differences in working memory and executive control: conceptual and methodological issues,” in *Handbook of Individual Differences in Cognition: Attention, Memory, and Executive Control*, eds A. Gruzka, G. Matthews, and B. Szymura (New York: Springer), 87–108.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 May 2011; paper pending published: 10 October 2011; accepted: 30 January 2012; published online: 05 March 2012.

Citation: Hartshorne JK and Schachner A (2012) Tracking replicability as a method of post-publication open evaluation. *Front. Comput. Neurosci.* 6:8. doi: 10.3389/fncom.2012.00008

Copyright © 2012 Hartshorne and Schachner. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



## APPENDIX

### SURVEY METHODS AND RESULTS

We contacted 100 colleagues directly as part of an anonymous Web-based survey. Colleagues of the authors from different institutions were invited to participate, as well as the entire faculty of one research university and one liberal arts college. Forty-nine individuals completed the survey: 26 faculty members, 9 post-docs, and 14 graduate students. Thirty-eight of these participants worked at national research universities. Respondents represented a wide range of sub-disciplines: clinical psychology (2), cognitive psychology (11), cognitive neuroscience (5), developmental psychology (10), social psychology (6), school psychology (2), and various inter-subdisciplinary areas.

The survey was presented using Google Forms. Participants filled out the survey at their leisure during a single session. The full text of the survey, along with summaries of the results, is included below. All research was approved by the Harvard University Committee on the Use of Human Subjects, and informed consent was obtained.

#### Part 1: Demographics

*Your research position: graduate student, post-doc, faculty, other* (26 faculty, 9 post-docs, and 14 graduate students).

*Your institution: national university, regional university, small liberal arts college, other* (38 national university, 4 regional university, 5 small liberal arts college, 2 other).

*Your subfield (cognitive, social, developmental, etc.; There is no standard set of subfields. Use your own favorite label):* \_\_\_\_\_

(11 cognitive psychology, 10 developmental psychology, 6 social psychology, 5 cognitive neuroscience, 2 school psychology, 2 clinical psychology, 13 multiple/other).

#### Part 2: completed replications

*In this section, you will be asked about your attempts to replicate published findings. When we say “replication,” we mean:*

*–a study in which the methods are designed to be as similar as possible to a previously published study. There may be minor differences in the method so long as they are not expected to matter under any existing theory. However, a study which uses a different method to make a similar or convergent theoretical point would be more than a replication. If you attempted to replicate the same finding several times, each attempt should be counted separately.*

*Given this definition...*

*1) Approximately how many times have you attempted to replicate a published study? Please count only completed attempts – that is, those with at least as many subjects as the original study.* \_\_\_\_\_

Total: 257; Mean: 6; Median: 2; SD: 11

(3 excluded: “NA,” “too many to count,” “50+”)

*2) How many of these attempts \*fully\* replicated the original findings?* \_\_\_\_\_

Excluding those excluded in (1):

Total: 127; Mean: 4; Median: 1; SD: 7

*3) How many of these attempts \*partially\* replicated the original findings?* \_\_\_\_\_

Excluding those excluded in (1):

Total: 77; Mean: 2; Median: 1; SD: 5

*4) How many of these attempts failed to replicate any of the original findings?* \_\_\_\_\_

Excluding those excluded in (1):

Total: 79; Mean: 2; Median: 1; SD: 4

*5) Please add any comments about this section here:* \_\_\_\_\_

[comments]

#### Part 3: aborted replications

*In this section, you will be asked about attempted replications that you did not complete (e.g., tested fewer participants than were tested in the original study).*

*1) Approximately how many times have you started an attempted replication but stopped before collecting data from a full sample of participants?* \_\_\_\_\_

Total: 48; Mean: 1; Median: 0; SD: 3

[3 excluded: “a few,” “countless,” (lengthy discussion)]

*2) Of these attempts, how many were stopped because the data thus far failed to replicate the original findings?* \_\_\_\_\_

Excluding those excluded in (1):

Total: 38; Mean: 2; Median: 0.5; SD = 4

*3) Of these attempts, how many were stopped for another reasons (please explain)?* \_\_\_\_\_

[comments]

*4) Please add any comments about this section here.*

[comments]

**Part 4: file-drawers**

1) *Approximately how many experiments have you completed (collected the full dataset) but, at this point, do not expect to publish?* \_\_\_\_

Total: 1312 (one participant reported “1000”); Mean: 31; Median: 3.5; SD: 154

(6 excluded: “many,” “ton,” “countless,” “30–50%?” 2 unreadable/corrupted responses)

2) *Of these, how many are not being published because they did not obtain any statistically significant findings (that is, they were null results)?* \_\_\_\_

Excluding those excluded in (1):

Total: 656 (one participant reported “500”); Mean: 17; Median: 2; SD: 81

3) *Please add any comments about this section here:* \_\_\_\_

[comments]



# Network-based statistics for a community driven transparent publication process

Jan Zimmermann<sup>1,2\*</sup>, Alard Roebroek<sup>1,2</sup>, Kamil Uludag<sup>1,2</sup>, Alexander T. Sack<sup>1,2</sup>, Elia Formisano<sup>1,2</sup>, Bernadette Jansma<sup>1,2</sup>, Peter De Weerd<sup>1,2</sup> and Rainer Goebel<sup>1,2,3\*</sup>

<sup>1</sup> Faculty of Psychology and Neuroscience, Department of Cognitive Neuroscience, Maastricht University, Maastricht, Netherlands

<sup>2</sup> Maastricht Brain Imaging Center (M-BIC), Maastricht University, Maastricht, Netherlands

<sup>3</sup> Department of Neuroimaging and Neuromodeling, Netherlands Institute for Neuroscience, an Institute of the Royal Netherlands Academy of Arts and Sciences (KNAW), Amsterdam, Netherlands

## Edited by:

Diana Deca, Technische Universität München, Germany

## Reviewed by:

Talis Bachmann, University of Tartu, Estonia

Rogier Kievit, University of Amsterdam, Netherlands

Dwight Kravitz, National Institutes of Health, USA

## \*Correspondence:

Jan Zimmermann and Rainer Goebel, Department of Cognitive Neuroscience, Maastricht University, Maastricht, Netherlands.  
e-mail: jan.zimmermann@maastrichtuniversity.nl;  
r.goebel@maastrichtuniversity.nl

The current publishing system with its merits and pitfalls is a mending topic for debate among scientists of various disciplines. Editors and reviewers alike, both face difficult decisions about the judgment of new scientific findings. Increasing interdisciplinary themes and rapidly changing dynamics in method development of each field make it difficult to be an “expert” with regard to all issues of a certain paper. Although unintended, it is likely that misunderstandings, human biases, and even outright mistakes can play an unfortunate role in final verdicts. We propose a new community-driven publication process that is based on network statistics to make the review, publication, and scientific evaluation process more transparent.

**Keywords:** network-based statistics, publishing system, scientific evaluation, peer review

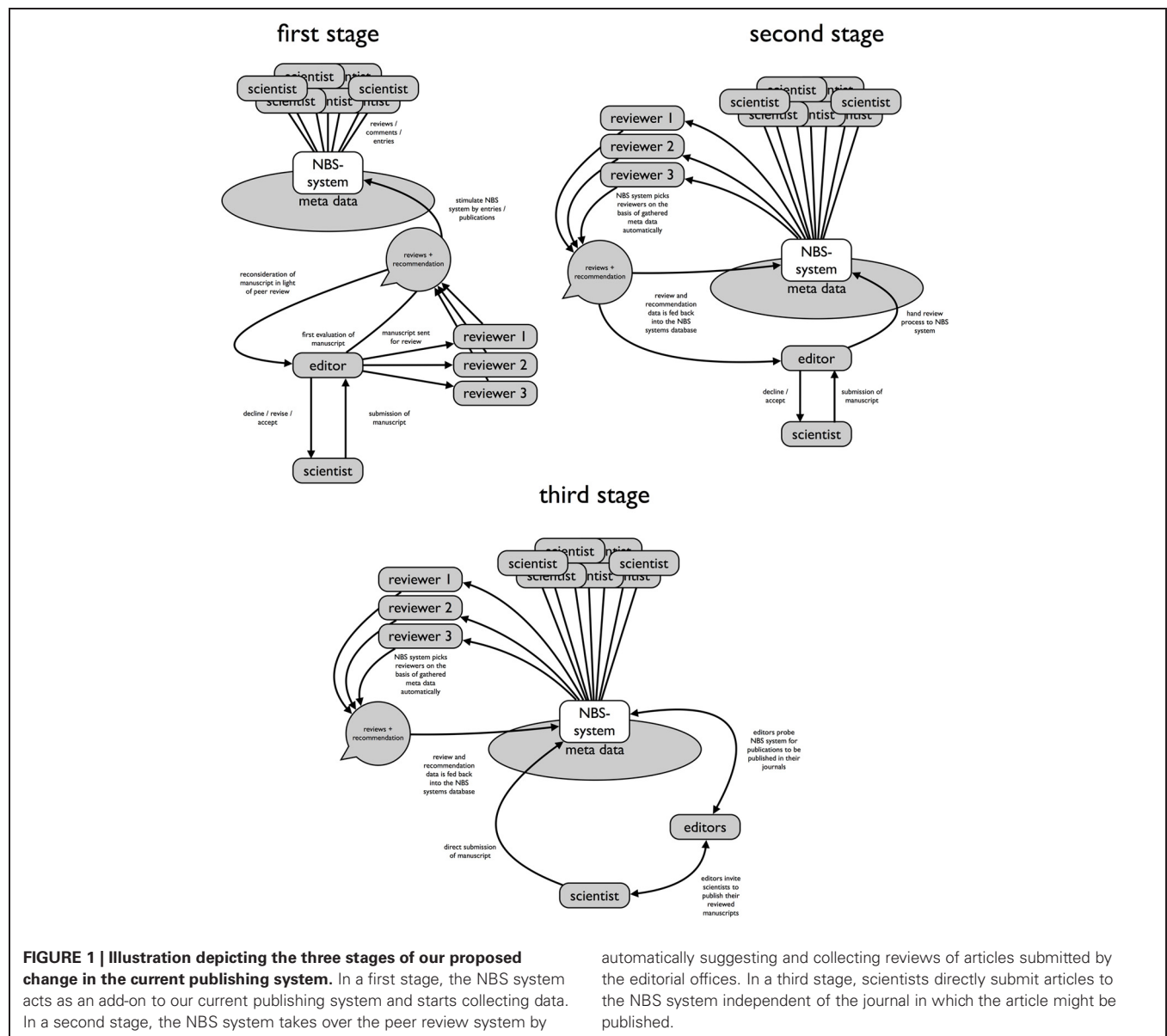
From an idealistic point of view, scientists aim to publish their work in order to communicate relevant findings. If we could rely on our own and individual judgment, review processes would not be needed. We obviously do not rely on our own judgment since more eyes see more and hence relevance and validity can be specified in a more objective way. Therefore, a system of peer review has been established as the method of choice to control for scientific relevance and methodological correctness/appropriateness. In fact, journal editors decide via the peer review process what is relevant and what in turn is communicated to other scientists via publication. Peer review has been the method of choice for many years, but scientists are concerned about the state of the current publishing system. Editorial as well as review decisions are not always fully transparent and vary between journals. The quality of a review depends on the expertise of the reviewer and the editorial office sometimes arbitrarily selects this expertise. The arbitrary element is a natural consequence of the task of the office and its realization in times of fast increase in submissions, the increase of interdisciplinary topics, and the lack of individual review expertise necessary to cover all issues of a modern science paper.

This discussion is not new at all. It has been stated before that the metrics by which the possible impact of an article is measured in the editorial handling phase are not well defined and leave a large degree of uncertainty about how decisions are made (Kreiman and Maunsell, 2011). The system is amenable to political as well as opportunistic biases playing a role in whether a paper is accepted or rejected (Akst, 2010). Public communication about an article and the review process to which it was subjected is very

limited, if possible at all. In addition, there is growing pressure from grant agencies and local institutions to publish a high number of articles, thereby potentially compromising the scientific quality of submitted papers, while the review process itself might be compromised by increased load due to the increasing number of submissions. Hence, we fear that the large increase in the number of publications in the field of neuroscience and other fields may be accompanied by a decrease in overall quality. Moreover, the explosion in numbers of publications makes it difficult to follow the evolution of a specific topic even for experts of that field. In the light of increasing financial pressure and importance of external funds, the reform of the publishing system cannot be viewed in isolation but has to take into account other parameters, which interact with the publishing system. Here, we provide an alternative to the current review and publishing system, which is meant to be implemented in two steps. The idea we propose is inspired by the development of social media. In the first step it would function as an add-on to the existing scientific publishing system, but in the second step may evolve to completely replace it. It involves the quantification of interactions among scientists using Network-Based Statistics (NBS), as done in social media, in combination with search tools, as used by Google. The proposal laid out below should act as an inspiration to where the future of publishing might lead, and is not intended to be a fully detailed roadmap.

## CURRENT STATE OF THE PUBLICATION PROCESS

In general, scientists submit an article covering their latest results and findings to a specific journal of interest (Figure 1, first stage,



bottom part). In most cases, a preliminary editorial decision is made whether the manuscript is of interest and of sufficient quality, after which the manuscript is either rejected or sent out for review to a small number of scientists (typically 2–3) who provide anonymous reviews of the submitted paper. The editor then faces a decision to accept the paper, to reject it, or to ask for revisions. This decision is to be guided by the Editor's own understanding of the topic, and the evaluation by the reviewers. If an article is rejected, the scientist may use the reviewers' concerns as a guideline to revising the manuscript for future submission in a different journal. If an article is accepted, the final version goes into the publishing stream of the journal and can be accessed by the community. In summary, the editorial and review decisions and the platform on which an article is presented, is tied to each individual journal and the accompanied publisher, and the process itself is usually entirely shielded from any public scrutiny.

## PROPOSED FUTURE STATE OF THE PUBLICATION PROCESS

We propose a new system that would initially accompany the existing one (**Figure 1**, first stage, top part), without generating excessive extra load for scientists and without increasing the already overwhelming number of published articles. The system would make use of modern technology to quantify the behavior of individuals in networks (NBS). The NBS system would initially function as an add-on to the existing system, but it might in a second stage lead to changes in the current system or to its replacement, by showing it is a superior system for all concerned. Evaluation of papers by NBS would be designed to be transparent and controlled by the scientific community. In short, the new system would quantify interactions among scientists pre- and post-publication, introduce new ways of determining an article's impact and, in a future stage, NBS would decouple the review process from individual journals and editors. The add-on NBS

system will work similarly to current social networks and would be built up of two types of general information; one being a scientific expertise profile of individual experts and the second being a database of publications (“entries”) with extended additional data (discussed below). Instead of maintaining scattered institutional websites containing individual information about publications, interests, and affiliations, scientists would subscribe to a global network where most important information about them is gathered. This information will include institutional affiliations, publications, and relationships to other collaborating scientists, which can be derived from author lists on publications and from statistical information about the behavior of scientists toward others (see below). Moreover, publications associated with member scientists would deliver information on the expertise and interests of each individual. Thus, the information provided can be used to extract metadata related both to expertise and connections of each individual in the network of scientists, and this information should be anonymously accessible by fellow scientists, editors, and publishers.

### **NBS AS A PARALLEL ADD-ON EXISTING NEXT TO THE CLASSICAL PUBLICATION PROCESS**

The proposed system can be used as an add-on to the current review system in the following way: when a new publication appears and when it is entered into the database (feeding of existing databases like Google scholar etc., or direct input by journals, thus having undergone traditional peer review), an editor associated with the NBS system will forward invitations to other scientists selected for their expertise and publication record to write brief comments, longer evaluations, or even extensive blog-like entries. This editor (or network administrator) will make the selection based on parameters provided by the NBS system, though the ultimate goal will be to generate the selection of reviewers and commentators on an automatic basis (see below). The quality and objectivity of a comment can be immediately evaluated, based on the metadata that is present in the system. For example, the position in the network relative to the authors on the publication can be objectively quantified in terms of numbers of common publications, overlap in (past and present) institutional affiliation, overlap in expertise, and content of previous comments (e.g., positive or negative), by algorithms accessing the metadata available in the system. Further statistical procedures could then be used (as in iTunes/Google) to find related comments, all entries from the same commenter, related entries from other commenters, etc. The combined results of such statistical data mining may greatly increase the transparency of evaluations and help scientists to weigh the importance of a paper versus its associated comments. In this initial stage, the NBS system, therefore, acts completely independent of the existing publishing and review system and adds an additional layer of information to each publication listed. This additional information provides an index to the reader about the relevance of a paper/topic within the community based on vividness of ongoing discussions about this paper. It is important to note, that the additional data should not act to replace the relevance, content, and substance as foundations of a given paper since those are not quantifiable in a direct way. However, the additional data can act in navigating

through the complex scientific landscape of publications where the final verdict on a paper should always be left to the critical scientific reader.

In addition, once the NBS system starts working, thus having gathered a sufficient amount of information, it may facilitate information clustering and career development. With regard to clustering, smart computer-driven clustering of comments in the database can be carried out in several dimensions (i.e., quality, quantity, type of author). They can then be used to visualize the relevance of a given paper over time. In addition to the comments left for a certain publication, usage of statistics such as views and downloads can be logged and taken into consideration during analysis of an articles history. This can be used as relevant orientation (and data reduction) for the scientific community and inherently contributes to scientific knowledge and quality. With regard to career development, the NBS can highlight competent and objective commentators on the basis of ratings and views. By doing so, NBS adds details to a scientist’s career profile in terms of impact (do people hear him/her) and vividness (quantity and quality of actions within NBS). NBS hence forms a tool to valorize scientific expertise via reviews as well as comments in general.

Taken together, the statistical information available can be used to provide measures that can promote more objective views on an article’s impact than its mere number of citations or the journals impact factor (Skorka, 2003; Simons, 2008; Franceschet, 2010), and provide a timeline of the importance it has on the scientific community. By having an ongoing assessment of a publication, clustering algorithms can be used to view a research topic and its related publications through the progression of time, independent from a single article’s reference list, even indicating what contributions individual manuscripts made to a specific domain of science. While substantive impact of a scientific idea is based on more than statistical data, the NBS system goes beyond the current standard metrics while making the process of judging impact more transparent. Proactive expertise contributions receive direct incentives as they are valued by the community. Since the NBS system relies on a large and valid amount of data, scientific institutions should support such proactive input by their scientists.

### **NBS AS AN ALTERNATIVE THAT CAN PARTLY OR COMPLETELY REPLACE THE EXISTING PUBLICATION PROCESS**

Initially, the NBS system would be based on the submission of papers that were published in journals, as well as unpublished papers, on which authors can comment in various formats similar to working papers which many disciplines are already familiar with. However, the network statistics associated with submitted articles and comments provide a parallel process that can be more than a mere add-on (**Figure 1**, second stage); we expect that the proposed system will be used to improve the current journal-driven reviewing system. Importantly, the system we propose with the scientist’s ability to comment on articles freely does not intend to replace the need for peer review in any way, only to restructure the process. Any manuscript submitted to the NBS system requires and should require a form of peer review, either directed by journals and their editors or by the system itself.



For example, the NBS system proposed here can be of immediate help to editors searching for relevant reviewers for a new article that has been submitted. A page rank algorithm, such as used by Google for retrieving information sorted by relevance to a keyword, could provide a relevant and, most importantly, scientifically objective reviewer to an editor. Objectivity could be defined as independent from the submitting scientists' group, affiliations, or personal preferences, but with overlapping expertise. Personal preferences and opportunistic behavior could be quantified based on an anonymous log of behavior among scientists. For example, scientists can be ranked by the tendency (quantified by appropriate metrics) to systematically reject papers of specific authors or institutions, and when this ranking index is too high, it should decrease their probability of being selected as a reviewer. By implementing such procedures, an editor using the proposed add-on system would enhance the review process by counteracting opportunistic behavior by individuals. While this system needs multiple occasions on which a reviewer is found to show this type of behavior, it is likely that its mere existence would reduce biases and make reviewers more aware of their claimed objectivity.

Furthermore, editors and scientists might agree to not only enter their papers into the NBS system, but also its anonymous reviews. Initially, this can be done with reviewers selected by a journal editor, who might have used the proposed system to select the reviewers. Importantly, at the discretion of the scientists authoring the paper and with permission of its reviewers, this would be done as soon as a paper has been reviewed, also if it is rejected. Each entry would, therefore, receive a history of its own review process prior to its ultimate publication in a journal. Hence, even if an article has not been accepted in a certain journal and ends up being published by another, the attached reviews should contain the entire publication process. Having the entire review process available for each article will make it more transparent for readers to judge how the reported findings were received as well as which problems (in terms of data acquisition, analysis methods, or hypothesis) fellow scientists tackled while getting published. Even for very good papers and positive reviews, an openly accessible review process might be enlightening as complementary additional ideas and background information would be shared (like a review of a good book or movie).

We believe that when editors start using this add-on system, it can influence journals and their editors to make better-informed decisions on how to select papers for publication. As our proposed NBS system would provide defined metrics of the success of an article, irrespective of where it gets published, or even whether or not it gets published, it would provide an alternative and more transparent measure of impact. We are convinced that NBS will provide more valuable measures of appreciation of a publication in a research field than classical impact measures and the journal's name. When editors increasingly use NBS to select reviewers, and when the view within the scientific field develops such that a system is beneficial, then consensus may grow. As a consequence, the current review process could be partly or entirely replaced by NBS. Indeed, it is imaginable in a third stage (**Figure 1**, third stage), that a system based on NBS would select reviewers for articles automatically based on objective statistics, and that what

initially would be comments would become the actual reviews of the submitted articles. In this way, a submitted article would generate its own review process that would be publicly available, in a way that is de-coupled from specific journals. Scientific journals would then be able to use the output of an NBS-based review processes to select articles for publication. This would create an inverse dynamic, in which journals will have to compete with each other to publish the best articles, as scientists might be contacted by several journals with requests for publication in print.

The scientific review and publication process we have sketched here will provide a context in which truly good publications will be labeled by favorable community-driven statistics and ranked high, while publications that were released prematurely or received poor ratings will also be recognized as such, and ranked low. We suggest that this will create a transparent and content-based competition among researchers and among institutions, so that quality of research may become emphasized more in evaluating an individual's productivity than numbers of publications. It can become a system that facilitates collaboration within the digitized social network. Moreover, we believe the proposed system will trigger a re-orientation of the effort of scientists from anonymous review processes that remain unpublished to interaction in a more open and public arena. We suggest that the more active and more publicly accessible communication style among scientists proposed here will lead to better knowledge of one another's work, and therefore, will be a catalytic factor in enhancing research quality.

## POSSIBLE CAVEATS AND DOWNFALLS

Any given system will have its inevitable flaws and problems and while we believe that our proposal aims at directly improving and addressing many of those present in our current systems state, it is important to note the possible problems our proposal could encounter. Scientific work, the content it entails and the quality associated with it is by its very nature not entirely quantifiable by metrics of statistics. Therefore, the proposed NBS system will never be independent of human evaluation instead we aim for making the system more transparent in that regard. It is clear that the system we propose has the possibility of generating excessive work load for scientists if mechanisms are not in place to control for endless discussion cycles. One serious problem with a more open system is the problem of danger of lobbyist tendencies. While opportunism and lobbyistism are problems already present in the current publishing system and we hope to alleviate them through means of the NBS system discussed above, it is important that activism within the NBS system does not counteract these efforts.

## SUMMARY AND OUTLOOK FOR THE FUTURE

Although it is difficult to predict how the introduction of the NBS-based publishing system would be received and thus develop, the minimal goal we wish to achieve is that publishers would increase the objectivity and transparency of the current review and publication system by using NBS-based information. This can be achieved by using NBS-based information for selecting reviewers, and scientists and editors agreeing to make the entire anonymized review history public on a publicly accessible

site (for a discussion on the problems associated with public reviews see Anderson, 1994 and Kravitz and Baker, 2011). However, in the long-term we suggest that a complete decoupling of the scientific review process from specific journals and from their different, idiosyncratic review systems would tremendously help the scientific objectivity of the review process. Indeed, scientific reviews should not be biased by the fact that a review is being handled for a high impact versus a lower impact journal, and it should not be biased by implicit histories or affinities of an author with a specific journal or editor. Moreover, a review system that is independent of individual editorial decisions and, therefore, not directly related to a particular journal would base the review process on a broader consensus-based evaluation.

Starting off our proposed add-on NBS-based system involves some, but minimal additional work by scientists (for a critical

view on electronic publications see Evans, 2008). It would involve an effort to make published articles accessible from a common webpage. Commenting/reviewing may be kicked-off by asking leading scientists to submit a number of comments on a subset of papers related to a topic of their competence. These comments will attract the scientific community to visit the system and to add further comments. This initial phase is essential in the development of the system in its add-on phase, and will be highly dependent on the effort of senior scientists. However, the overall benefits and possibilities of the new system should cover these initial costs entirely. We strongly believe that it is time to leave the sub-optimal reviewing and publication system that is available right now behind, and reform it into a more transparent and open system. Importantly, to make this transition effective, universities, research organizations, and grant agencies have to be part of the reform and support it.

## REFERENCES

- Akst, J. (2010). I hate your paper. *Scientist* 24, 36–41.
- Anderson, R. H. (1994). Anonymity of reviewers. *Cardiovasc. Res.* 28, 1735.
- Evans, J. A. (2008). Electronic publication and the narrowing of science and scholarship. *Science* 321, 395–399.
- Franceschet, M. (2010). Journal influence factors. *J. Informetrics* 4, 239–248.
- Kravitz, D. J., and Baker, C. I. (2011). Toward a new model of scientific publishing: discussion and a proposal. *Front. Comput. Neurosci.* 5:55. doi: 10.3389/fncom.2011.00055
- Kreiman, G., and Maunsell, J. H. R. (2011). Nine criteria for a measure of scientific output. *Front. Comput. Neurosci.* 5:48. doi: 10.3389/fncom.2011.00048
- Simons, K. (2008). The misused impact factor. *Science* 322, 165.
- Skorka, P. (2003). How do impact factors related to the real world? *Nature* 425, 661.
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Received: 24 November 2011; paper pending published: 27 December 2011; accepted: 17 February 2012; published online: 05 March 2012.
- Citation: Zimmermann J, Roebroek A, Uludag K, Sack AT, Formisano E, Jansma B, De Weerd P and Goebel R (2012) Network-based statistics for a community driven transparent publication process. *Front. Comput. Neurosci.* 6:11. doi: 10.3389/fncom.2012.00011
- Copyright © 2012 Zimmermann, Roebroek, Uludag, Sack, Formisano, Jansma, De Weerd and Goebel. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Letting the daylight in: reviewing the reviewers and other ways to maximize transparency in science

Jelte M. Wicherts\*, Rogier A. Kievit, Marjan Bakker and Denny Borsboom

Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

## Edited by:

Nikolaus Kriegeskorte, Medical  
Research Council Cognition and Brain  
Sciences Unit, UK

## Reviewed by:

Nikolaus Kriegeskorte, Medical  
Research Council Cognition and  
Brain Sciences Unit, UK  
Tal Yarkoni, University of Colorado at  
Boulder, USA

Jasper Jacobus Franciscus Van Den  
Bosch, Goethe-Universität Frankfurt  
am Main, Germany

## \*Correspondence:

Jelte M. Wicherts, Department of  
Psychology, University of Amsterdam,  
Weesperplein 4, 1018 XA Amsterdam,  
Netherlands. e-mail:  
j.m.wicherts@uva.nl

With the emergence of online publishing, opportunities to maximize transparency of scientific research have grown considerably. However, these possibilities are still only marginally used. We argue for the implementation of (1) peer-reviewed peer review, (2) transparent editorial hierarchies, and (3) online data publication. First, peer-reviewed peer review entails a community-wide review system in which reviews are published online and rated by peers. This ensures accountability of reviewers, thereby increasing academic quality of reviews. Second, reviewers who write many highly regarded reviews may move to higher editorial positions. Third, online publication of data ensures the possibility of independent verification of inferential claims in published papers. This counters statistical errors and overly positive reporting of statistical results. We illustrate the benefits of these strategies by discussing an example in which the classical publication system has gone awry, namely controversial IQ research. We argue that this case would have likely been avoided using more transparent publication practices. We argue that the proposed system leads to better reviews, meritocratic editorial hierarchies, and a higher degree of replicability of statistical analyses.

**Keywords:** peer review, scientific policy, data sharing, scientific integrity

## INTRODUCTION

It has been argued, most famously by Karl Popper, that the openness of the scientific system is what makes it such a successful epistemic project, compared to other methods of gathering knowledge. The open character of scientific arguments allows the error-checking mechanisms of science, such as replication research, to work. In turn, this eradicates incorrect claims efficiently so that, in science, falsehoods tend to die young. It seems safe to say that openness is so central to the value system of the scientific community, that occasions where we choose *not* to pursue an open system should be as rare as possible. In principle, such occasions should only arise when there are overriding concerns of a higher moral status, such as concerns with regard to the privacy of patients participating in research and similar factors. From this point of view, it is remarkable that one of the most important parts of the scientific process, peer review, takes place behind closed curtains.

This hidden part of science has some undesirable consequences. For instance, it means that essential parts of the scientific discussion are invisible to the general audience. In addition, the peer review system is liable to manipulation by reviewers and editors. For example, editors can influence the system by selecting subsets of reviewers who, given their track record, are practically certain to provide positive or negative reviews. Reviewers can manipulate the system by “bombing” papers; especially top journals tend to publish papers only if all reviewers judge a paper positively, so that a single dissenting vote can nip a submission in the bud.

These and other problems with the peer review system have been widely debated (e.g., Godlee et al., 1998; Smith, 2006; Benos

et al., 2007), yet the system has been subject to little change. One reason may be that the peer review system is a case where we are both “us” and “them”: practicing scientists both bear the adverse consequences of its problems and are responsible for its faults. Moreover, the editorial secrecy itself precludes the reviewing scandals that occur from becoming public and creating sufficient outrage to provide adequate momentum for change. A final problem is that scientists have grown accustomed to the system; so even though many see it as a wicked labyrinth, at least it is one in which they know how to navigate.

So general are the problems of the peer review system and so (seemingly) hard to remedy that some have likened peer review to democracy, in being “a bad system, but the best we have” (e.g., Moxham and Anderson, 1992; Van Raan, 1996). However, as is the case for democracy, the fact that peer review is both inherently imperfect (as is any human endeavor) and likely to remain at the heart of scientific publishing does not imply it cannot be improved. In fact, we will suggest a simple improvement that may go a long way toward solving the current problems; namely, to open up the peer review system itself. In this context, we will propose a new system that is based on three pillars: (1) the publication of reviews, (2) the public assessment of the quality of those reviews, and (3) mandatory publication of data together with a published paper.

We argue that this system has several immediate payoffs. First, it is likely to improve the overall quality of reviews, especially by allowing the scientific community to discount reviews that are clearly biased or which provide too little argumentation. Second, the system remedies the lack of direct acknowledgment of the work that goes into reviewing, which is a significant drawback

of the current system, and one of the primary reasons that it is becoming harder for editors to find reviewers. Third, making the system public opens up further insights into the structure of the scientific literature. Compared to current practices in scientific publishing, the proposed system is based more strongly on the key characteristics of the scientific enterprise: honesty, openness, and rigor. As we illustrate in the next sections, current practice of reviewing and dealing with research data do not always do well in these regards.

We will delve more deeply into a specific example, but first note that cases of controversial peer review decisions exist in most if not all fields of science. In the last 2 years alone, there have been several examples of high-profile research where peer review has, seemingly, not functioned well. For instance, *Science* accepted for publication a paper by Wolfe-Simon et al. (2011) that claimed to have found evidence for arsenic-based life forms, thereby overturning basic assumptions in (molecular) biology. However, colleagues heavily criticized the paper almost instantly, with several very critical commentaries appearing (e.g., Redfield, 2010). The paper was eventually published along with eight highly critical comments and an editorial note (Alberts, 2011). Similarly, *Nature* published a paper by influential theorists that argued that kin selection is an outdated concept (Nowak et al., 2010). The paper immediately sparked controversy, and was followed in a later volume of the same journal by several critical replies, one of which had 136 authors (Abbott et al., 2011). Arguably the most damaging case of peer review gone awry was an article by Wakefield et al. (1998) in *The Lancet*, allegedly demonstrating a link between vaccines and autism. The article, based on 12 patients, was ultimately retracted, the lead author's medical license revoked, and the claims stricken from the academic record after an intensive investigation revealed several cases of fraud. Although fraud cannot always be detected by peer review, inspection revealed several grave errors such as improper measures, lack of disclosure of conflicting interests, improper blinding procedures and a lack of controls that could have been picked up by peer review (for an overview, see Godlee et al., 2011).

The breadth of the critique in these controversial cases, generally representing the majority of scientists in the respective fields, lends credence to the hypothesis that the reviewing process was, at the very least, not as rigorous as is desirable. Several controversial examples make clear that poorly reviewed papers, given the current dearth of opportunity to correct such errors, can adversely affect progress of science and in some cases (i.e., the Wakefield paper) be damaging to the public. As science's main method of quality control, it is clear that all parties would benefit from a peer review system that diminishes the chances of such errors occurring.

We will illustrate the nature of the problems with current peer review and our proposed solution on the basis of a case that, in our view, represents the problems with the current system most clearly. As the variety of examples above show, this particular case is not of great importance. We chose it because (a) we are familiar with its content and the context in which it appeared, (b) we feel confident in judging the merits of the paper and the problems that should have been picked up by reviewers, and (c) its problems *could have* been solved in a more open system of peer review. If we

succeed in our goal, readers will be able to substitute our particular case study with a relevant example from their field.

## A CASE STUDY

### THE CASE

On the basis of his theory of the evolution of intelligence (Kanazawa, 2004), Kanazawa (2008) proposed that, during their evolutionary travels away from the relatively stable and hence predictable environment of evolutionary adaptedness (EEA; i.e., the African savanna of the late Pleistocene), the ancestors of Eurasians encountered evolutionarily novel environments that selected for higher intelligence. Therefore, Kanazawa (2008) predicted higher average IQ scores in countries located farther away from the EEA. Kanazawa (2008) tested this hypothesis against data gathered by Lynn and Vanhanen (2006), who estimated so-called "national IQ-scores," i.e., the average IQ of the inhabitants of nations in terms of western norms. Kanazawa (2008) found a significant negative correlation between countries' national IQs and their distance from three geographic locations in and around sub-Saharan Africa.

### WHAT SHOULD HAVE HAPPENED?

We point to a number of indisputable issues that should have precluded publication of the paper as constituted at the time of review. First, Kanazawa's (2008) computations of geographic distance used Pythagoras' theorem and so the paper assumed that the earth is flat (Gelade, 2008). Second, these computations imply that ancestors of indigenous populations of, say, South America traveled direct routes across the Atlantic rather than via Eurasia and the Bering Strait. This assumption contradicts the received view on evolutionary population genetics and the main theme of the book (Oppenheimer, 2004) that was cited by Kanazawa (2008) in support of the Out-of-Africa theory. Third, the study is based on the assumption that the IQ of current-day Australians, North Americans, and South Americans is representative of that of the genetically unrelated indigenous populations that inhabited these continents 10,000 years ago (Wicherts et al., 2010b). In related work by others who share Kanazawa's (2008) views on the nature of race differences in IQ, the latter issue was dealt with by excluding countries with predominantly non-indigenous populations (Templer and Arikawa, 2006). Thus, although Wicherts et al. (2010b) raised additional issues that may the topic of debate (see below), these three problems are beyond dispute.

### WHAT DID HAPPEN?

The paper was accepted for publication in the journal *Intelligence* 3 weeks after first submission. *Intelligence* is the foremost journal on human intelligence and has an impact factor of 3.2<sup>1</sup>. The editor normally asks three experts to review original and revised submissions. Editorial decisions concerning rejection, acceptance, or revision are based on the majority vote, although one critical reviewer may be sufficient to let authors revise the manuscript several times. The average time lag for research papers that were published in 2008 was 228 days (median = 211) and so the acceptance of Kanazawa's (2008) paper was rapid.

<sup>1</sup>One of us (Jelte M. Wicherts) is proud to be a member of its editorial board although he hastens to add he was not one of Kanazawa's (2008) reviewers.



## AFTERMATH

Two of the authors of the present paper were involved in the preparation of a criticism that pointed out some of the undisputable errors in the paper, and also raised doubts with respect to the evidential relevance of present day correlations for evolutionary theories of the kind Kanazawa (2004, 2008) proposed. After we had submitted the critique to *Intelligence* we received the following feedback from two anonymous reviewers. According to Reviewer 1 of our critique: “The history of science tells us that a strong theory that explains numerous phenomena, like that of [...] Kanazawa, is generally overturned by a better theory, rather than by the wholly negative and nitpicking criticisms of the present paper.” Reviewer 2 of our comment wrote that: “Any explanation of IQ biodiversity must address itself to the totality of the evidence and not depend on highlighting small scale criticisms.” A third reviewer was more positive, but the use of the majority vote resulted in rejection of our criticism.

## ANALYSIS OF THE CASE

Because we have no access to the reviews of Kanazawa's (2008) paper, we can only speculate on how the review process unfolded. Having a clear bearing on the controversial topic of race differences in IQ one would expect Kanazawa's (2008) study to be met with scrutiny by reviewers (Hunt and Carlson, 2007). This does not appear to have happened. It is possible that the reviewers were busy and each hoped for other reviewers to scrutinize the paper in detail. In psychology, such processes have been studied in detail under the headers of *social loafing* and *diffusion of responsibility* (Darley and Latané, 1968), and are known to negatively influence the quality of task performance.

Another possibility is that Kanazawa's (2008) reviewers performed poorly because they felt the need to counter the unpopularity of views associated with genetic hypotheses of group differences in IQ. Our view is that the current state of knowledge of the neurophysiological, evolutionary, genetic, cognitive, and psychometric nature of individual differences in IQ is insufficient to arrive at clear answers about the nature of group differences in IQ. However, the topic is certainly a legitimate scientific endeavor, and we take no issue with researchers who propose hypotheses that feature racial differences in genetic endowment for intelligence (as long as these hypotheses are testable and consistent). Yet many researchers consider those who hypothesize on such genetic differences to be racist and not even entitled to publish their work in a peer-reviewed journal. Dishonest reviews in this controversial area are well documented on both sides of the debate (Hunt, 1999; Gottfredson, 2010). Dishonest reviews are the atrocities in the “wars of science” and their existence only sparks more dishonesty, which does not really contribute to knowledge.

## AN ALTERNATIVE HISTORY

The fate of our critique of Kanazawa's (2008) paper (and of two similar papers by others) is interesting, because it provides an alternative history by itself. The reason is that the journal *Personality and Individual Differences* eventually published the paper, along with a polite and open debate (Lynn, 2010; Rushton, 2010; Temple, 2010; Wicherts et al., 2010a,b) on the relevance of some of the additional issues we had raised earlier (unfortunately Kanazawa

himself declined the invitation to comment). The exchange clearly shows that opinions on Kanazawa's (2008) findings differ. The differences in tone and content between the negative reviews of our earlier manuscript and the open exchange in the other journal are striking. One likely reason is that the reviews were written anonymously and in a system that is not sufficiently open to scrutiny. Although editors play a moderating role in debates between authors and reviewers (next to their main role in deciding on publication), they are unlikely to disagree with reviewers for several reasons. First, editors need to be able to fall back on the reviewers' assessments to make unpopular rejection decisions and to be able to counter later criticisms of published work. Second, the editors rely on these reviewers in the future to do more pro bono reviewing. Similarly, it is impolite to ask busy scientists to invest time to review a paper and subsequently downplay or ignore the importance of their work. Writing peer reviews takes up valuable time but these writings are normally not published and so the editors are unlikely to complain when the reviews are done hastily.

## CONCLUSION

In our view the case study illustrates a major problem with current publication practices. Namely that the selection of reviewers, editorial decision making, and the treatment of critiques are all done behind closed curtains and that reviewers are often anonymous, and so hardly accountable for their writings. The general audience may thus read the paper in *Intelligence* without recognizing that it is based on several faulty assumptions, and without ever knowing that a criticism of the paper was rejected. Nor can the audience ever retrace the arguments that led to the acceptance of Kanazawa's (2008) paper and rejection of the criticism voiced against it. The general audience has no way of finding out how three reviewers who are knowledgeable in their field had missed the publication of obvious errors they were supposed to help avoid and how two reviewers later prevented an exposition of these errors in the same outlet. Peer reviews represent some of the most valuable and interesting reflections on other peoples' work and putting them away in a closed system is often a waste of energy and information. Also, the payoffs for reviewers to write high quality reviews are currently minor.

Let us then consider a new system, based on the premise of complete openness, discuss its possible merits and drawbacks, and finally examine a brief counterfactual history of the case study to illustrate how the peer reviewing system might work, and why this is a benefit for all concerned.

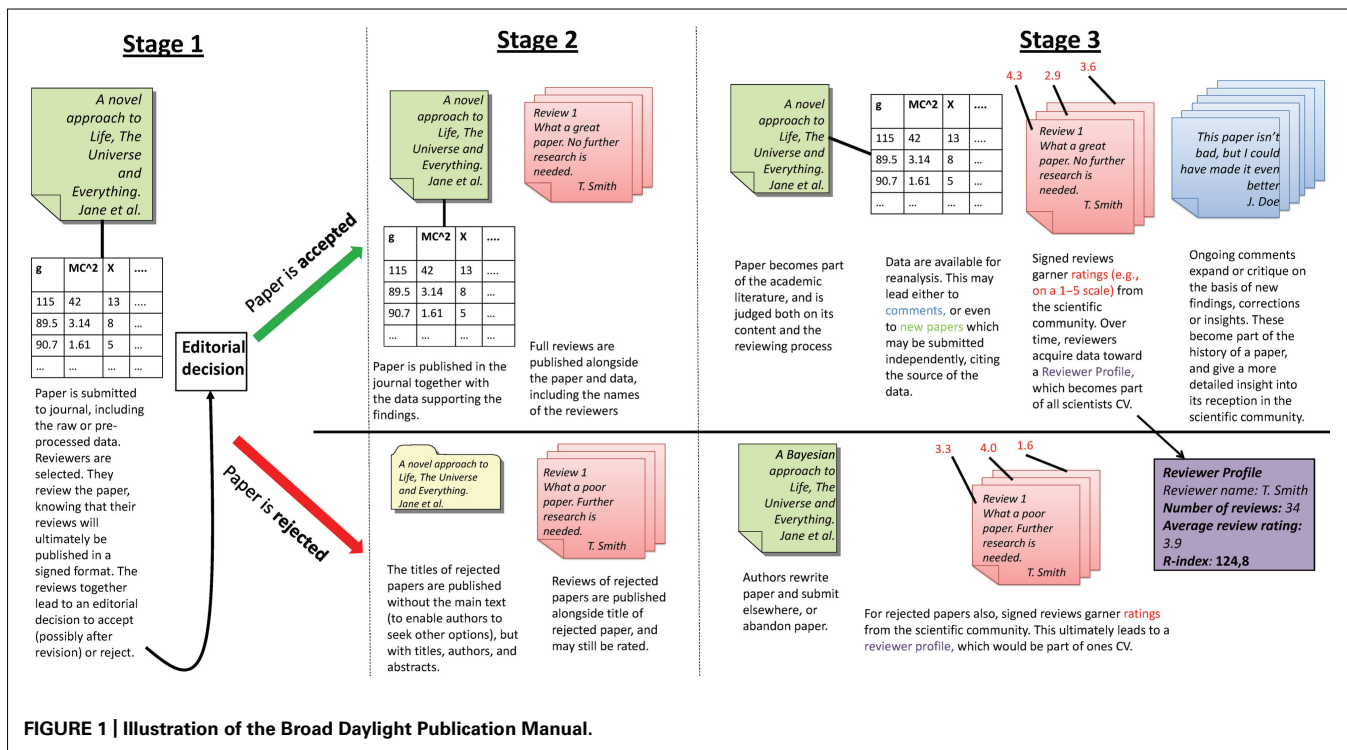
## THE BROAD DAYLIGHT PUBLICATION MODEL

Fortunately, there is an effective cure for all of these diseases: daylight. The Broad Daylight Publication Model (BDPM) that we advance here incorporates openness at three levels: transparency of the editorial process, accountability of reviewers, and openness with respect to data. The BDPM is illustrated in **Figure 1**.

## THE EDITORIAL PROCESS

The BDPM first involves a soft change to current policy. It merely requires giving up secrecy and opening up the scientific system as it exists now to public scrutiny. This means that scientific journals should disclose all information by default, unless there are





overriding concerns to preclude such practice. So journals should minimally engage in the following steps:

1. **Disclose submissions:** All submissions should eventually be published online, so that the public may see not only which papers were accepted, but also which papers were rejected. Rejected papers are published without the main text (to enable authors to seek other options), but with titles, authors, and abstracts, and full reviews.
2. **Disclose reviews:** All reviews of all papers, whether accepted or rejected, should eventually be published online, along with all editorial letters.

We think that the current secrecy regarding who submitted what where and how the submission was evaluated is outdated. Only rarely do authors have insurmountable reasons to remain secret about their submitted work. Almost certainly, reviewers would write their reviews differently if they knew that these reviews will become public.

### ACCOUNTABILITY OF REVIEWERS

A second step in promoting openness involves making reviewers accountable for their actions and to give them due credit for their hard reviewing work. This could be done by adding the following elements:

3. **Review the reviewers:** All reviews of all papers can be rated by the journal's readership. Reviews are always signed.
4. **Open up the editorial hierarchy:** Reviewers who review often and whose reviews get high ratings can ascend in the editorial hierarchy.

We propose a system where every review can itself be rated by the scientific community. We suggest some criterion that warrants

the ability to be able to rate reviews, such as “having at least one published article in this journal.” Any person who fulfills this criterion may then rate a review on a Likert scale that runs from, say, 1 to 5. These ratings represent the perceived quality, depth, expertise of the review, and the extent to which it contributes to quality control. After publishing the reviews alongside the manuscript, these reviews will accumulate ratings. After some time, a review may have scored an average of, say “4.2,” suggesting fairly high average review quality. Similarly, a reviewer will start accruing ratings and published reviews. We could think of some basic metric (e.g., for instance an “R-index,” that summarizes “number of reviews written” times “average quality rating”) that reflects both the amount and average quality of reviews someone has conducted, which would be a relevant part of the resume of a working scientist. This would allow the work that goes into reviewing to be acknowledged more explicitly, and for funding agencies to judge someone’s “presence” in the scientific community more accurately. In this way, reviewing well will finally start to pay off for the reviewers themselves. By writing many reviews that are published alongside manuscripts, researchers may build their reputation in the community. A good reputation as a reviewer should form the basis for appointments in the editorial hierarchy (reviewing board, editorial board, associate editors, and main editor).

Another important effect of opening up the review system is that structure of the reviewing process can be analyzed. For instance, it would become possible to examine patterns of friendly reviewing and nepotism. In addition, reviewers can be statistically analyzed. It will be clear to everyone living in the scientific machine that reviewers differ in how difficult it is to pass them. Such differences can be analyzed and, in the future, it may even be possible to account for them. In fact, the availability of such

information enables a wealth of studies that contribute to scientific self-reflection and improve the scientific practice, thereby advancing knowledge.

Importantly, it is also possible to see who gave which ratings, and if there are large discrepancies. All parties will benefit from highly rated reviews: the authors of the original manuscript as their paper has withstood high quality scrutiny, the reviewers themselves because their contribution has been acknowledged and reported upon, possibly leading to editorial promotion, and the journal and its editor as they have, in the perception of the larger community, succeeded in appointing appropriate reviewers. Altogether, peer review of reviews will improve the quality of the published work. We also feel that this will improve the quality of reviews of rejected papers toward being more constructive.

Another benefit is that the quality of the journal may be assessed also by the reviewing standards it sets. The impact factor of a journal is commonly used as the predominant indicator of its quality. However, we could easily envisage a situation where a journal increases in stature for the overall quality of the reviews upon which it bases its decisions. This average rating would represent the expertise, fairness, and scientific judgment of the editor. This would be especially relevant for journals that are highly specialized and therefore generally have a low impact factor, such as *Psychometrika* in our own field. This journal has low citation statistics, but is highly regarded by both applied and theoretically oriented psychometricians for its rigor and high quality standards. The rating of the reviews may offer such journals a new metric, on which the community can base its judgment: one that reflects the rigor and quality of its reviewing standards, and therefore the presumed quality of its academic content, not just the popularity of the articles it publishes. Journals with many highly regarded reviews are also expected to receive more submissions.

As is the case for papers (in which other theories are often critiqued), people should be accountable for their assessment of a paper. Currently, scientists are quite comfortable praising or discrediting theories or techniques within the confines of their own papers and/or commentaries, so there should be no reason why people will suddenly refuse to critique (or compliment) work openly in reviews. Ultimately, it is the editor who makes the decision; the reviewers merely give a recommendation.

Consistently writing highly regarded reviews, regardless of the decisions that they lead to, could and should be used as the basis of appointing editors of journals. A reputation for rigorous and fair reviews is probably not easily earned, and should be rewarded. Published reviews could be considered publications in their own right. Currently, commentaries are considered to be separate publications, even though they are shorter than conventional manuscripts.

## OPENING UP THE DATA

Finally, as the BDPM requires opening up the scientific system, not only the submissions and reviews should be disclosed, but the data should be published as well. Although the ethical guidelines of for example the American Psychological Association (2010) require data sharing on request, the current practice holds that data are *not* shared unless exceptional circumstances hold (Wicherts et al., 2006; Savage and Vickers, 2009). The right policy is clearly to

publish the empirical data on which empirical claims are based, *unless exceptional circumstances apply* (e.g., privacy issues, data ownership). Thus, we argue that the research data of studies should be submitted to the journal as a matter of scientific principle as soon as a paper is accepted for publication (Wicherts and Bakker, 2012), which leads to our fifth principle:

5. *Disclose the data:* Data should be published online along with the papers whose empirical claims they support.

Several practical issues need to be dealt with. First, the confidentiality of the human participants needs to be protected. This can be dealt with in several ways. Data can be anonymized and release of particular data can be restricted to those who can be held responsible for protecting the confidentiality. Exemption can be requested when data are overly sensitive or when legal issues preclude the release of proprietary data. Second, researchers who collected the data may wish to conduct future research with the data after the first results are published. This problem can be dealt with at the researchers' request by imposing, say, an 18-month moratorium on the release of the data (or a moratorium proportional to the cost of acquiring a given dataset). This should give the original researchers a reasonable head start on their competition. Third, data require proper documentation. Fortunately, there are several successful data archives in numerous fields of science. Quality standards of data archiving are well developed (e.g., see <http://www.datasealofapproval.org/>). However, it is of importance to develop guidelines on documenting and archiving neuroscientific data, which present specific challenges.

Considering data as an integral part of any publication has been proposed by many, including Hanson et al. (2011, p. 649) in a recent editorial in *Science*: "As gatekeepers to publication, journals clearly have an important part to play in making data publicly and permanently available." Although research data lie at the core of science, they are normally published only in highly condensed form as the outcomes of the statistical analyses that the researcher happened to report. Quite often the raw data can tell us considerably more than a single *p*-value, or a single brain image showing pooled differential activity. Specifically, researchers may disagree on how the data should be analyzed, new analyses may provide new insights on the findings, and independent re-analyses of the data may expose errors in the statistical analyses (Wicherts and Bakker, 2012).

Straightforward checks on the basis of basic information in papers show an alarmingly high prevalence of statistical errors, even in the most prestigious journals (Rossi, 1987; Garcia-Berthou and Alcaraz, 2004; Murphy, 2004; Berle and Starcevic, 2007; Strasak et al., 2007; Kriegeskorte et al., 2009; Nieuwenhuis et al., 2011). For instance, after a simple check of the consistency between reported test statistics and *p*-values in a fairly representative sample of 257 papers published in psychology, Bakker and Wicherts (2011) found that nearly half of these papers contained at least one error in the reporting of statistical results. In roughly one in seven papers they found a result that was unjustly reported as being significant. In another study it was found that researchers who report such erroneous results are less likely to share their data for reanalysis (Wicherts et al., 2011). As these errors were identifiable from just the information present in the published studies,

they could have been prevented by sound statistical review. By making reviews both public and accountable, more errors might get identified (e.g., because spotting of such errors is likely to be a straightforward way to gain a high profile as a statistical reviewer.) However, these errors might just be the tip of the iceberg. Other statistical errors can only be exposed with access to the raw data. In addition, availability of the raw data may help prevent scientific misconduct (Wicherts, 2011).

Apart from statistical errors, the details of statistical analyses typically affect what can be concluded from the data. Results are often dependent on decisions like how to transform the data, the methods used in averaging across subjects or over time, or the identification of outliers. Analyzing neuroscientific data in particular can be a complex task in which statistical decision making may lead to published effects that appear to be inflated (Kriegeskorte et al., 2009; Vul et al., 2009). On top of that, researchers often have a lot to gain in finding and being able to report an interesting (and often significant) result. Since in many scientific fields (with the notable exception of some medical fields; ICH, 1996) statistical choices are not explicated in advance in statistical protocols, the researcher often has a lot of room to maneuver in doing the analyses. The fact that many actually do capitalize on this freedom is evidenced by the statistically unlikely (Sellke et al., 2001) overrepresentation of *p*-values just below the typical 0.05 threshold for significance that has been documented in various fields that involve traditional data analyses (Ioannidis and Trikalinos, 2007; Ridley et al., 2007; Gerber and Malhotra, 2008a,b). If contention exists about the decisions and analyses, the only scientific way to resolve the issue is to have the raw (or pre-processed) data available for anyone to examine. At the end of the day, whether such re-analyses should be considered nitpicking or pertinent to the hypothesis of the paper is to be judged by the scientific community.

Of course, data sharing will not only serve as a quality control device (although this is a crucial aspect). There are many positive incentives for the scientific community. One of those clear benefits is the more efficient (re)use of existing data. Especially in fields that rely on complex, computationally heavy analyses such as behavior genetics, (cognitive) neuroscience, and global climate models, sharing data will vastly increase the availability of data to validate new techniques and uncover previously unnoticed empirical phenomena in existing data. Examples of successful data sharing programs are the Human Genome Project<sup>2</sup>, Neurosynth<sup>3</sup>, and the BrainMap Project<sup>4</sup>. Data that have already been published could be used for additional studies without much additional cost. Reusing data will perhaps shift the focus away from “new data” (several high-impact journals explicitly state that data should not have been published before) and toward new *findings*.

## THE FATE OF A PAPER IN THE BDPM

Given the above, what would happen if one submitted a paper in the broad daylight paper system? A paper is submitted to the desired journal, including the dataset (stripped of any identifiers and pre-processed if necessary) on which the conclusions

are based. This paper, including the dataset, is sent to a selection of reviewers with the necessary expertise. After an appropriate timeframe, they submit their reviews and the recommendations (reject, revise and resubmit, accept) that follow from their reviews. The editor then decides on the basis of these reviews whether or not to accept the paper, possibly weighing the reviews on the basis of previous reviewer quality ratings (i.e., one of the reviewers may have a high average rating for his or her previous reviews). If the paper is ultimately published, it is published on the website of the journal. The website contains the manuscript, the editorial decision, the reviews, and the raw (or pre-processed) data. Colleagues can then, after reading the manuscript and the reviews, rate those reviews on a scale of 1–5 (based on the “at least one publication” rule). These ratings represent a guide for a new reader of the manuscript, both to its virtues and possible problematic components. Finally, readers may comment on the manuscript and so review the paper themselves after it has been published. Although such later reviews play no role in decisions concerning acceptance of the paper, they do allow the community to comment on it. Like the original reviews, these later comments entail a manner to make a career as a reviewer/commenter. After a period of time, this would create a dynamic representation of the validity and quality of the paper. Does it stand up to scrutiny? Are the reviews upon which publication was based considered to be rigorous? Are any potential flaws pointed out in the later comments? Let us now re-examine the Kanazawa (2008) case from the perspective of this new system, and how this is an improvement.

## A COUNTERFACTUAL HISTORY OF THE CASE STUDY

What then, in our system, would have become of the case study, and why is it an improvement? The paper would have been submitted to the same journal. We consider it quite likely that it would have been met with more criticism and that the indisputable errors discussed above would have been averted in earlier phases of the review. Perhaps reviewers would have opposed publication, but let us suppose that they would have recommended publication. Subsequently, the paper and its reviews would have become available for all to read. If the system works as we envisage it, several things that we consider an improvement could happen.

Firstly, anyone (including journalists) will be able to read the paper, but also the reviews on which the acceptance was based, the ratings these reviews received, and whether they were sufficiently critical. This will go a long way in judging whether to accept the (possibly controversial) views put forth. On the basis of this assessment, people may then rate those reviews in terms of thoroughness, scientific credibility, and general quality. We expect many readers of *Intelligence* to not have rated the reviews of Kanazawa's (2008) paper highly.

Secondly, readers may comment informally (and under their own names) on the paper as much as is currently possible in journals like *PLoS ONE*. This would allow for instantaneous feedback, both positive and negative, on the merits and possible flaws of the manuscript and its reviews. Currently, it is no exaggeration to state that the impact factor of the journal is often considered the most important factor in judging the merits of an individual paper. This is clearly a rather crude heuristic, better replaced by discussions and feedback on the actual paper.

<sup>2</sup>[http://www.ornl.gov/sci/techresources/Human\\_Genome/home.shtml](http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml)

<sup>3</sup><http://neurosynth.org>

<sup>4</sup>[www.brainmap.org](http://www.brainmap.org)

Finally, readers may use the data (that was made available alongside the manuscript) to evaluate the data, to consider alternative hypotheses and perhaps to even be inspired to re-analyze the data in a way that provides even more, or different, support of the theory under consideration. Unlike many other instances in which data are unavailable after publication (Wicherts et al., 2006, 2011), Kanazawa's (2008) data could be submitted to secondary analyses. These analyses cast some doubt on his hypotheses (Wicherts et al., 2010b; Hassal and Sherratt, 2011).

Over time, this would lead to a changing and dynamic consideration of the merits of the paper, based on the quality of the reviews (as judged by readers), the general tone of comments and whether or not any convincing counterarguments are put forth over time, possibly based on new analyses. Or, of course, someone might find a fatal flaw. Notably, the converse may also be the case: if all the negative comments are only based on ideological critiques, and not on substantive or scientific arguments, this may be considered implicit support for the claims in the paper, regardless of their (un)popularity. Of course, the best possible scenario is that the open nature and dynamics of the BDPM create a community where there are clear incentives for thorough reviewing. We hope that all readers would consider this alternative history to be preferable over what actually happened in this specific case.

## FEASIBILITY OF THE BDPM

One could argue that our system may sound good in theory, but that the reality of incentives and the sociological dynamics of science are such that they are not compatible with a fully open system. We think that although this has some superficial plausibility, a closer inspection of specific problems shows that none are insurmountable, and that these problems are outweighed by its benefits.

## OPENNESS

Will people be willing to review openly? Although the fear that people will not be willing to sign their reviews openly seems reasonable, empirically, this does not seem to be the case. Smith (2009) has an interesting empirical finding: "Interestingly, when we asked a sample of reviewers whether they would review openly about half said yes and half no. When we conducted the trial, very few people declined to review openly and when we introduced the policy only a handful of reviewers in a database of around 5,000 refused to sign reviews." Medical journals published by BioMed Central have successfully introduced a system in which signed reviews are published alongside the published papers. Although Godlee et al. (1998) did not find clear benefits of having reviewers sign their reviews, such benefits may well appear when the reviews are published and subsequently rated by readers.

## HONESTY

Will people be equally honest? Another fear may be that the visibility of reviews will lead people to sugarcoat their reviews, where they would have criticized sub-par work more harshly in the past. One plausible fear may be the imbalance of power in the community. For instance, a young and upcoming researcher may not want to make any enemies, thus "pulling punches." This may be

the case, but we cannot envisage this to be a big problem. Even a cursory glance at the literature shows that scientists are generally not reluctant to criticize one another. In fact, in our view it is far more likely that the scientific community appreciates honest, well-founded critique, regardless of whether someone is a scientific veteran or a starting graduate student. And if someone does tend to pull his or her punches, this will become apparent in the BDPM as overly tame signed reviews from this person accumulate. An "accept as is" from someone who is also occasionally critical and regularly rejects papers may be more valuable than an "accept as is" from someone who always recommends publication.

## PARTICIPATION

A glance at some of the existing online possibilities of post-publication commenting (e.g., at *PLoS ONE*) shows that not all papers will be heavily commented on. Perhaps not all reviews will be rated. This is not a problem of the new system, but a simple fact concerning the sheer volume of scientific production. Not all papers will be widely read, not all papers will be cited, and not all papers will have a large impact. This already applies to even the highest impact-journals (e.g., Mayor, 2010). The greatest benefit of the BDPM is that it offers the tools and opportunities for correction, falsification and quality control, and gives increased insight into the background of a paper. Moreover, by introducing a system in which the ratings of reviews have an influence on the selection of reviewers and even editorial positions, we expect a stronger involvement by the community.

## ABUSE

Some may fear that a reward system based on ratings is easily exploitable. However, given that users can view ratings by name, we think the simple fact of having traceable ratings will largely diminish this problem. Everyone can see where the ratings of the reviews come from. This may serve to expose an excessive degree of nepotism. Although it is perfectly natural (and highly likely) that people rate the work of their colleagues highly, insight into who gave which votes will again allow people to judge what they think of a manuscript. If, say, all the people with a statistics background rate a review poorly, that may be an incentive to partly discount a review that argues that inappropriate analyses were used.

## LOGISTICAL ISSUES IN DATA SHARING

Although data files from many studies in the medical and behavioral sciences are quite straightforward and are readily archived, this does not apply to most multidimensional data files from neuroscience. There is a clear need for guidelines and best practices of the sharing of such complex data files. The extensive pre-processing of neuro-imaging data should be documented in ways that enable replication on the basis of the raw data, whereas pre-processed data that were used in the published analyses could be submitted to the journal. Rigorous documentation of data handling and the archiving of the raw data (even if these data are submitted to more specialized repositories or simply stored at the academic institution) is essential for replication and is required by ethical guidelines. Major funding organizations increasingly demand that data are shared (Wicherts and Bakker, 2012) and so



the costs associated with sharing of data should become an integral part of research funding. We are aware of previous failed attempts of journals (like the *Journal of Cognitive Neuroscience* in the mid 1990s) to implement policies of data sharing, but we feel that the times are changing. As the number of (high-impact) journals with such policies increases so will researchers' willingness to share.

## CONCLUSION

In sum, we do not see insurmountable problems in setting up a truly open scientific publication system. Our moral principle of openness as a default mode of science, rather than as an exception, thus suggests that we should simply start implementing such a system. Increased transparency at various levels would, in our view, eradicate a number of practices that arise under the current shroud

of secrecy. Editorial manipulation through choice of reviewers would be exposed almost immediately. Low quality and/or biased reviews would, in our view, quickly disappear under the pressure of daylight. Accepting papers that include gross errors would certainly become more difficult. Due to the possibility of earning credits through good reviewing, reviewing itself would finally start to pay off. Data would become publicly accessible, which not only allows for replicating the statistical analyses, but also archives the data for use by future generations of scientists. There is no system without drawbacks. However, all things considered the proposed ways of increasing transparency appear desirable. It remains to be seen how researchers react to increased openness; it is entirely possible that they will happily embrace it. There is only one way to find out: just do it.

## REFERENCES

- Abbott, P., Abe, J., Alcock, J., Alizon, S., Alpedrinha, J. A. C., Andersson, M., Andre, J. B., van Baalen, M., Balloux, F., Balshine, S., Barton, N., Beukeboom, L. W., Biernaskie, J. M., Bilde, T., Borgia, G., Breed, M., Brown, S., Bshary, R., Buckling, A., Burley, N. T., Burton-Chellew, M. N., Cant, M. A., Chapuisat, M., Charnov, E. L., Clutton-Brock, T., Cockburn, A., Cole, B. J., Colegrave, N., Cosmides, L., Couzin, I. D., Coyne, J. A., Creel, S., Crespi, B., Curry, R. L., Dall, S. R., Day, T., Dickinson, J. L., Dugatkin, L. A., El Mouden, C., Emlen, S. T., Evans, J., Ferriere, R., Field, J., Foitzik, S., Foster, K., Foster, W. A., Fox, C. W., Gadau, J., Gandon, S., Gardner, A., Gardner, M. G., Getty, T., Goodisman, M. A., Grafen, A., Grosberg, R., Grozinger, C. M., Gouyon, P. H., Gwynne, D., Harvey, P. H., Hatchwell, B. J., Heinze, J., Helanterä, H., Helms, K. R., Hill, K., Jiricny, N., Johnstone, R. A., Kacelnik, A., Kiers, E. T., Kokko, H., Komdeur, J., Korb, J., Kronauer, D., Kümmerli, R., Lehmann, L., Linksvayer, T. A., Lion, S., Lyon, B., Marshall, J. A., McElreath, R., Michalakakis, Y., Michod, R. E., Mock, D., Monnin, T., Montgomerie, R., Moore, A. J., Mueller, U. G., Noë, R., Okasha, S., Pamilo, P., Parker, G. A., Pedersen, J. S., Pen, I., Pfennig, D., Queller, D. C., Rankin, D. J., Reece, S. E., Reeve, H. K., Reuter, M., Roberts, G., Robson, S. K., Roze, D., Rousset, F., Rueppell, O., Sachs, J. L., Santorelli, L., Schmid-Hempel, P., Schwarz, M. P., Scott-Phillips, T., Shellmann-Sherman, J., Sherman, P. W., Shuker, D. M., Smith, J., Spagna, J. C., Strassmann, B., Suarez, A. V., Sundström, L., Taborsky, M., Taylor, P., Thompson, G., Tooby, J., Tsutsui, N. D., Tsuji, K., Turillazzi, S., Ubeda, F., Vargo, E. L., Voelkl, B., Wenseleers, T., West, S. A., West-Eberhard, M. J., Westneat, D. F., Wiernasz, D. C., Wild, G., Wrangham, R., Young, A. J., Zeh, D. W., Zeh, J. A., and Zink, A. (2011). Inclusive fitness theory and eusociality. *Nature* 471, E1–E5.
- Alberts, B. (2011). Editor's note. *Science* 332, 1149.
- American Psychological Association. (2010). *Publication Manual of the American Psychological Association*, 6th Edn. Washington, DC: American Psychological Association.
- Bakker, M., and Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behav. Res. Methods* 43, 666–678.
- Benos, D. J., Bashari, E., Chaves, J. M., Gagar, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., McGowan, S., Polter, A., Qadri, Y., Sarfare, S., Schultz, K., Splitterger, R., Stephenson, J., Tower, C., Walton, R. G., and Zotov, A. (2007). The ups and downs of peer review. *Adv. Physiol. Educ.* 31, 145–152.
- Berle, D., and Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *Int. J. Methods Psychiatr. Res.* 16, 202–207.
- Darley, J. M., and Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *J. Pers. Soc. Psychol.* 8, 377–383.
- Garcia-Berthou, E., and Alcaraz, C. (2004). Incongruence between test statistics and P values in medical papers. *BMC Med. Res. Methodol.* 4, 13. doi: 10.1186/1471-2288-4-13
- Gelade, G. A. (2008). The geography of IQ. *Intelligence* 36, 495–501.
- Gerber, A. S., and Malhotra, N. (2008a). Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Q. J. Polit. Sci.* 3, 313–326.
- Gerber, A. S., and Malhotra, N. (2008b). Publication bias in empirical sociological research – do arbitrary significance levels distort published results? *Sociol. Methods Res.* 37, 3–30.
- Godlee, F., Gale, C. R., and Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports a randomized controlled trial. *JAMA* 280, 237–240.
- Godlee, F., Smith, J., and Marcovitch, H. (2011). Wakefield's article linking MMR vaccine and autism was fraudulent. *Br. Med. J.* 342, 64–66.
- Gottfredson, L. S. (2010). Lessons in academic freedom as lived experience. *Pers. Individ. Dif.* 49, 272–280.
- Hanson, B., Sugden, A., and Alberts, B. (2011). Making data maximally available. *Science* 331, 649.
- Hassal, C., and Sherratt, T. (2011). Statistical inference and spatial patterns in correlates of IQ. *Intelligence* 39, 303–310.
- Hunt, E. B., and Carlson, J. S. (2007). Considerations relating to the study of group differences in intelligence. *Perspect. Psychol. Sci.* 2, 194–213.
- Hunt, M. (1999). *The New Know-nothings: The Political Foes of the Scientific Study of Human Nature*. New Brunswick, NJ: Transaction Publishers.
- ICH. (1996). *Good Clinical Practice: Consolidated Guidance*. Geneva: International Conference on Harmonisation.
- Ioannidis, J. P. A., and Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clin. Trials* 4, 245–253.
- Kanazawa, S. (2004). General intelligence as a domain-specific adaptation. *Psychol. Rev.* 111, 512–523.
- Kanazawa, S. (2008). Temperature and evolutionary novelty as forces behind the evolution of general intelligence. *Intelligence* 36, 99–108.
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12, 535–540.
- Lynn, R. (2010). Consistency of race differences in intelligence over millennia: a comment on Wicherts, Borsboom and Dolan. *Pers. Individ. Dif.* 48, 100–101.
- Lynn, R., and Vanhanen, T. (2006). *IQ and Global Inequality*. Augusta, GA: Washington Summit Publishers.
- Mayor, J. (2010). Are scientists near-sighted gamblers? The misleading nature of impact factors. *Front. Psychol.* 1:215. doi: 10.3389/fpsyg.2010.00215
- Moxham, H., and Anderson, J. (1992). Peer review: a view from the inside. *Sci. Technol. Policy* 5, 7–15.
- Murphy, J. R. (2004). Statistical errors in immunologic research. *J. Allergy Clin. Immunol.* 114, 1259–1264.
- Nieuwenhuis, S., Forstmann, B. U., and Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.* 14, 1105–1107.
- Nowak, M. A., Tarnita, C. E., and Wilson, E. O. (2010). The evolution of eusociality. *Nature* 466, 1057–1062.
- Oppenheimer, S. (2004). *Out of Eden: The Peopling of the World*. London: Constable & Robinson Ltd.
- Redfield, R. (2010). Arsenic-associated bacteria (NASA's claims). Retrieved on April 15, 2011, from <http://rrresearch.blogspot.com/2010/12/arsenic-associated-bacteria-nasas.html>
- Ridley, J., Kolm, N., Freckleton, R. P., and Gage, M. J. G. (2007). An unexpected influence of widely used significance thresholds on the distribution of reported P-values. *J. Evol. Biol.* 20, 1082–1089.
- Rossi, J. S. (1987). How often are our statistics wrong – a statistics class exercise. *Teach. Psychol.* 14, 98–101.
- Rushton, J. P. (2010). Brain size as an explanation of national differences in IQ, longevity, and other life-history variables. *Pers. Individ. Dif.* 48, 97–99.



- Savage, C. J., and Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE* 4, e7078. doi: 10.1371/journal.pone.0007078
- Sellke, T., Bayarri, M. J., and Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Am. Stat.* 55, 62–71.
- Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *J. R. Soc. Med.* 99, 178–182.
- Smith, R. W. (2009). In search of an optimal peer review system. *J. Particip. Med.* 1, e13.
- Strasak, A. M., Zaman, Q., Marinell, G., Pfeiffer, K. P., and Ulmer, H. (2007). The use of statistics in medical research: a comparison of The New England Journal of Medicine and Nature Medicine. *Am. Stat.* 61, 47–55.
- Templer, D. I. (2010). Can't see the forest because of the trees. *Pers. Individ. Dif.* 48, 102–103.
- Templer, D. I., and Arikawa, H. (2006). Temperature, skin color, per capita income, and IQ: an international perspective. *Intelligence* 34, 121–139.
- Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics* 36, 397–420.
- Vul, E., Harris, C., Winkelman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4, 274–290.
- Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., Berelowitz, M., Dhillon, A. P., Thomson, M. A., Harvey, P., Valentine, A., Davies, S. E., and Walker-Smith, J. A. (1998). Retracted: ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet* 351, 637–641.
- Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature* 480, 7.
- Wicherts, J. M., and Bakker, M. (2012). Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence* 40, 73–76. doi: 10.1016/j.intell.2012.01.004
- Wicherts, J. M., Bakker, M., and Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE* 6, e26828. doi: 10.1371/journal.pone.0026828
- Wicherts, J. M., Borsboom, D., and Dolan, C. V. (2010a). Evolution, brain size, and the national IQ of peoples around 3,000 years B.C. *Pers. Individ. Dif.* 48, 104–106.
- Wicherts, J. M., Borsboom, D., and Dolan, C. V. (2010b). Why national IQs do not support evolutionary theories of intelligence. *Pers. Individ. Dif.* 48, 91–96.
- Wicherts, J. M., Borsboom, D., Kats, J., and Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *Am. Psychol.* 61, 726–728.
- Wolfe-Simon, F., Blum, J. S., Kulp, T. R., Gordon, G. W., Hoeft, S. E., Pett-Ridge, J., Stolz, J. F., Webb, S. M., Weber, P. K., Davies, P. C., Anbar, A. D., and Oremland, R. S. (2011). A bacterium that can grow by using arsenic instead of phosphorus. *Science* 332, 1163–1166.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 May 2011; accepted: 16 March 2012; published online: 03 April 2012.

Citation: Wicherts JM, Kievit RA, Bakker M and Borsboom D (2012) Letting the daylight in: reviewing the reviewers and other ways to maximize transparency in science. *Front. Comput. Neurosci.* 6:20. doi: 10.3389/fncom.2012.00020

Copyright © 2012 Wicherts, Kievit, Bakker and Borsboom. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# Decoupling the scholarly journal

Jason Priem\* and Bradley M. Hemminger

School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

## Edited by:

Nikolaus Kriegeskorte, Medical  
Research Council Cognition and  
Brain Sciences Unit, UK

## Reviewed by:

Thomas Boraud, Université de  
Bordeaux, France  
Nikolaus Kriegeskorte, Medical  
Research Council Cognition and  
Brain Sciences Unit, UK

## \*Correspondence:

Jason Priem, School of Information  
and Library Science, University of  
North Carolina at Chapel Hill, 100  
Manning Hall, Chapel Hill, NC  
27599-3360, USA.  
e-mail: priem@email.unc.edu

Although many observers have advocated the reform of the scholarly publishing system, improvements to functions like peer review have been adopted sluggishly. We argue that this is due to the tight coupling of the journal system: the system's essential functions of archiving, registration, dissemination, and certification are bundled together and siloed into tens of thousands of individual journals. This tight coupling makes it difficult to change any one aspect of the system, choking out innovation. We suggest that the solution is the "decoupled journal (DcJ)." In this system, the functions are unbundled and performed as services, able to compete for patronage and evolve in response to the market. For instance, a scholar might deposit an article in her institutional repository, have it copyedited and typeset by one company, indexed for search by several others, self-marketed over her own social networks, and peer reviewed by one or more stamping agencies that connect her paper to external reviewers. The DcJ brings publishing out of its current seventeenth-century paradigm, and creates a Web-like environment of loosely joined pieces—a marketplace of tools that, like the Web, evolves quickly in response to new technologies and users' needs. Importantly, this system is able to evolve from the current one, requiring only the continued development of bolt-on services external to the journal, particularly for peer review.

**Keywords:** scholarly communication, peer review, publishing, models

## INTRODUCTION

Why have we failed to reform peer review? It is certainly not for lack of trying; the last few decades have seen growing awareness of the institution's glaring weaknesses, and a plethora of alternatives suggested. We suggest that there are two reasons reform has been lacking:

1. Changes to peer review are just patches on a fundamentally broken scholarly journal system.
2. Proposals offer no smooth transitions from the present system.

In this paper, we suggest a reform of peer review that is built atop a reform of the entire publishing system. Importantly, though, we also argue that this new system can evolve in incremental steps, each viable on its own, from the present one. To guide us, we borrow the idea of "refactoring."

Refactoring is a programming practice in which we look at a computer system, identify parts that are confusing, inefficient, or redundant, and then systematically improve them—all while making sure that the functions of the program do not change (Hendler, 2007; Ding et al., 2009). We propose a refactoring of the scholarly journal system. This starts with an analysis of the current system, which we will do in the next section. We then proceed to suggest a better system, the "decoupled journal (DcJ)." After reviewing similar solutions proposed by others, we describe the DcJ in detail, and give some examples of what it would look like in practice. We close by considering advantages of our proposal, particularly how scholars can smoothly transition to it from the current model.

## THE CURRENT SCHOLARLY JOURNAL SYSTEM FUNCTIONS OF THE JOURNAL

Our first step in analyzing the scholarly journal system is to determine its functions. These are our constraints: whatever we change about the system, we must make sure that it can still perform these functions. Next we examine how the functions are currently being performed—the structure of the system. Finally we look for ways in which the current structure seems inefficient or redundant, and propose improvements

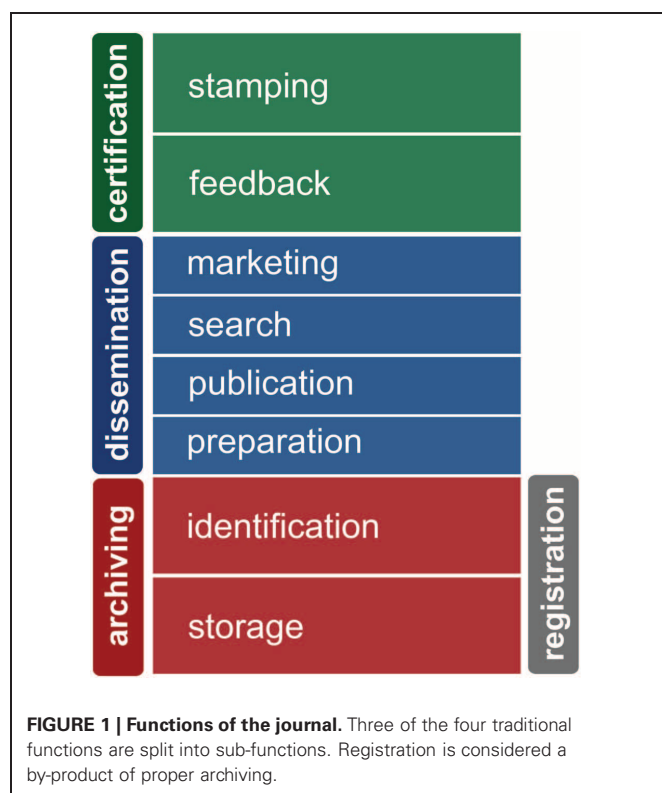
An authoritative list of functions is well beyond our scope. However, over the decades a consensus has emerged in the literature that journals have four "traditional functions" (Rowland, 2002):

- Archiving: permanently storing scholarship for later access.
- Registration: time-stamping authors' contributions to establish precedence.
- Dissemination: getting scholarly products out to scholars who want to read them.
- Certification: assessing contributions and giving "stamps of approval."

Over the years many authors have suggested additional or alternate functions (many are listed in **Table 1**). We will base our analysis on the traditional functions, since they are as close to an authoritative list as is available. However, observing that several proposed functions seem to be sub-functions of the traditional four, we incorporate them as well. We also add a few observed sub-functions of our own, finally giving us the more detailed

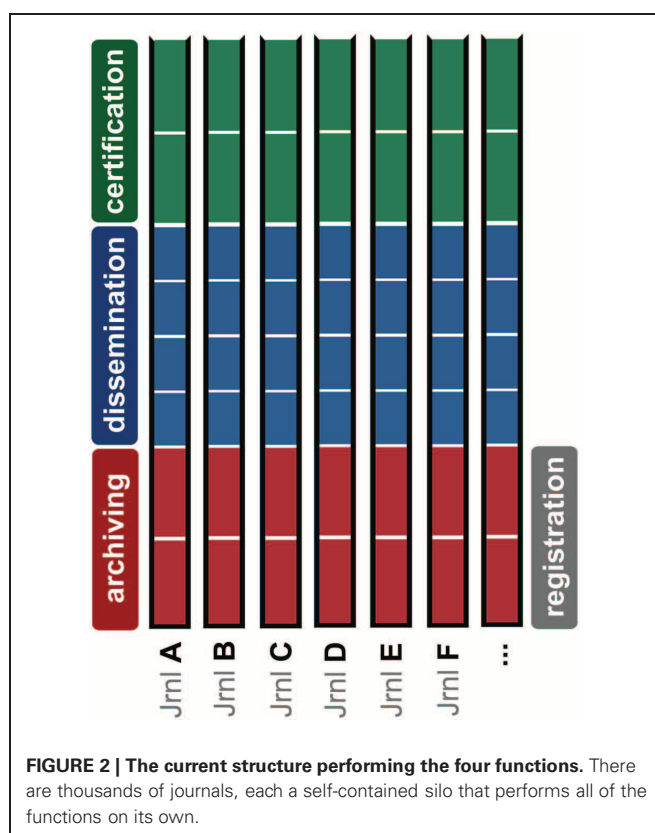
**Table 1 | Functions of the journal outside the traditional four.**

Filtration	Clarke (2010)
Rewarding	Clarke (2010) and Roosendaal and Geurts (1997)
Marketing	Smith (2003)
Cataloging	Smith (1999)
Copypediting	Rowland (2002) and Smith (1999)
Defining subject and community	Donovan (1998) and Smith (1999)
Democracy	Hendler (2007)
Retrieval	Casati et al. (2007)



model of the journal's functions show in **Figure 1**. This model honors the consensus around the traditional four functions, while at the same time allowing us to examine the diverse functions of the journal in greater detail.

We note that certification, for example, does not just consist of giving out seals of approval to worthy work—the feedback that authors get from reviews is also a valuable function. Dissemination has the greatest number of sub-functions; it requires some form of manuscript preparation (copypediting and typesetting), marketing, and provision for search, in addition to the actual publication. Archiving necessitates both persistent storage and identification. We break a bit with tradition by collapsing the registration function with archiving, as it seems clear that any system meeting the needs of the latter will fulfill the function of the former as well. Likewise, we omit proposed functions like “rewarding” that are *supported* by the journal system, but not



actually one of its functions (the reward proper comes from one's peers, university or granting agency).

### STRUCTURE OF THE JOURNAL

Our next step is to examine the structure of the journal system to see how well it supports the performance of its functions. Again, a full-scale analysis, such as Ware and Mabe's (2009) is well out of our scope. However, three particularly maladaptive features of the current structure are readily apparent:

1. The market is split into around 25,000 individual journals (Ware and Mabe, 2009), each one performing *all four functions* more or less in isolation as seen in **Figure 2** (van de Sompel et al., 2004).
2. The business model is dominated by the selling of content to readers, and consequently tends to value secrecy and closedness.
3. Peer review, the lynchpin of the entire system, shows remarkably little variation or innovation in practice—despite a troubling opacity, observed bias (Peters and Ceci, 1982; Wenneras and Wold, 2008), inefficiency, and lack of empirical support (Jefferson et al., 2007).

The last two of these problems have seen sustained and high-profile attention from policy-makers, thought leaders, and a growing percentage of the academy's rank and file. However, we argue that while it might not be apparent at first glance, the first problem is actually the most serious, and in fact leads to the other two.

The Balkanization of the scholarly literature was not planned; indeed, the journal was supposed to be a cure for just this problem. Oldenburg, creator of the first scientific journal in 1665, realized that instead of mailing letters to one another, as was the contemporary practice, scientists could communicate more efficiently by mailing to a central location and then disseminating all the letters together. Scholars today still care much less about the journal than what it contains, and this sense grows as they increasingly access literature through vast, flat indexes like PubMed and Google Scholar. Ultimately, there is just one journal: the scholarly literature (Gordon and Poulin, 2008), a conceptual space we dub the metajournal.

The persistent fragmentation of the metajournal leads to appalling diseconomies of scale. Perversely in this age of ever-growing academic specialization, we have a system of journals that are still technical generalists—an archipelago of self-sufficient islands, each blithely performing all four functions in splendid isolation. Journals as they now exist are jacks of all the communication trades, but consequently masters of none.

More seriously, the bundling together of all the functions in a single entity has stifled innovation by making it hard to experiment with individual functions—like peer review, or open access—without the expense and risk of creating whole new journals. I can choose a journal to publish in or read, but I cannot in most cases ask the journal for a particular kind of review. Bundling the functions together insulates any one function from the market, allowing poor implementations to flourish and preventing good ones from being directly rewarded. This explains the slow change in business models and peer review models that have perplexed many forward-thinking academics and publishers (Greaves et al., 2006; Gotzsche et al., 2010; Schriger et al., 2011). We suggest that no amount of activism or innovation aiming to correct closed publishing models or broken certification models will succeed in the current system that closely bundles all the functions together.

There is a good analogy here to another concept in programming: that of *separation of concerns* (SoC) (Reade, 1989). Concerns are the different sorts of things a program does: presenting output, receiving commands, communicating over the network, and so on. The idea is that if each of the concerns is handled in relative isolation from the others, it is much easier to maintain, repair, and improve its handling, because doing so doesn't disturb the rest of the system. If, on the other hand, SoC is violated, improving a single feature requires modifying or even rewriting the entire system. This is exactly the current problem facing scholarly communication: the journal system has fused the functions together in such a way that consumers have little choice regarding individual functions, and innovators must tackle the entire system in order to change a few pieces.

## THE DECOUPLED JOURNAL: PROPOSITION AND HISTORY

### THE DECOUPLED JOURNAL

Borrowing another programming term, we suggest that any solution to the problems of publishing must start with *decoupling* (Stevens et al., 1974). In software this means making the pieces of the systems as small, distinct, and modular as possible. This can be done for the journal, as well. We know the functions of

the scholarly journal. Let's make communication services that pick just one of those functions; then, do it well. The basic providers of scholarly publishing should not be publishers or journals, but smaller, more specialized, more modular services. This will let us assess different segments' performance more clearly, spot inefficiencies more quickly, and correct them more easily. The central virtue of a DcJ is, as in the case of a decoupled program, the system's ability to adapt to change quickly and relatively painlessly, because any given piece is as can easily be replaced.

To use a metaphor outside computing, the current journal system is like a fixed-price menu, in which a few sets of courses are selected for diners in advance, and ordered as one item. This has the benefit of simplicity. But its inflexibility means that diners don't get to exercise their creativity, and the chef may never realize that the risotto isn't any good—you just can't get the quail without it. We advocate scholarly communication à la carte—letting diners combine courses as they please so they get the meal that is most satisfying at the best price. Our goal is not to change the functions of the journal, but to remix (or rather, un-mix) them, taking advantage of profound technological change in the centuries since the system was developed.

### PREVIOUS IMPLEMENTATIONS

There are several publishing paradigms that *partly* decouple the journal, and these deserve a closer look. Of these, we will examine overlay journals, PLoS One, post-publication review services, and Smith's (1999) proposed Deconstructed Journal.

#### Overlay journals

Overlay journals, as first suggested by Ginsparg (1997) are journals that *only* perform the certification function; they peer review material already published, archived, and registered in an external repository, and publish a simple link for each accepted article (Moyle and Lewis, 2008; Brown, 2010). Repositories can be institutional repositories (IRs) or subject-area repositories like the *ArXiv*.

There have been several interesting prototypes of tools for creating and managing overlay journals, as well as a number of function examples in the wild. The RIOJA project (Moyle and Lewis, 2008) created an overlay journal system based on Open Journal Systems, a popular application for managing open access journals (Willinsky, 2003). Also in the UK, the *Overlay Journal Infrastructure for the Meteorological Sciences* project created a demo overlay journal. Rodriguez et al. (2006) created an interesting prototype of an overlay journal system that uses the co-authorship graph to automatically select appropriate reviewers for articles in distributed repositories, then adds review information as metadata using *The Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH). In addition to these demonstration projects, **Table 2** lists examples of real journals that have actually implemented the overlay model.

Overlay journals are promising because they could allow experimentation in peer review and other functions without the burden of managing entirely new journals. By offloading responsibility for archiving, dissemination, and registration to external repositories, overlay journals demonstrate that scholarly journals

**Table 2 | Overlay journals in the wild.**

Journal name	Journal URL	Listed as overlay in	Currently overlay (hosting articles on repository server)
Journal of High Energy Physics	<a href="http://www.springer.com/physics/particle+and+nuclear+physics/journal/13130">http://www.springer.com/physics/particle+and+nuclear+physics/journal/13130</a>	Brown (2010)	No
Logical Methods in Computer Science	<a href="http://www.lmcs-online.org">http://www.lmcs-online.org</a>	Brown (2010)	Yes
Geometry and Topology	<a href="http://www.msp.warwick.ac.uk/gt">http://www.msp.warwick.ac.uk/gt</a>	Brown (2010), UC Davis Front for the Archive list at <a href="http://front.math.ucdavis.edu/journals">http://front.math.ucdavis.edu/journals</a>	No
Journal of Nonlinear Mathematical Physics	<a href="http://staff.www.ltu.se/~norbert/home_journal/">http://staff.www.ltu.se/~norbert/home_journal/</a>	Front for the ArXiv	No
Algebraic and Geometric Topology	<a href="http://www.msp.warwick.ac.uk/a">http://www.msp.warwick.ac.uk/a</a>	Front for the ArXiv	No
Advances in Theoretical and Mathematical Physics	<a href="http://www.intlpress.com/ATM">http://www.intlpress.com/ATM</a>	Front for the ArXiv	No

can indeed be decoupled and still succeed. However, the history of the overlay journal is not particularly encouraging for proponents of decoupling. The idea has existed some time, and has apparently failed to ignite the imaginations of potential publishers. Indeed, all but one of the journals in **Table 2** have abandoned the overlay model and returned to traditional, highly coupled publishing.

What accounts for this disappointing reaction? It is impossible to know for sure, and it would be interesting to pursue research asking editors of journals who had switched to traditional publishing their reasons. However, one reason might be technical: until recently, the available tools were optimized for traditional journals; simply archiving and publishing authors' work as a conventional journal may have been easier than managing an overlay infrastructure, especially given the low cost of electronic storage. However, perhaps a deeper problem is that overlay journals do not pursue the decoupling idea *far enough*: they split the roles of the journal it two, but perhaps it needs to be split yet further.

### **PLoS One**

Another approach partly decoupling the journal comes from the journal *PLoS One*. This is an unconventional title that publishes work from any scientific discipline, provides free access for readers, and uses a relatively novel approach to peer review: reviewers are *specifically told* not to consider a work's significance or potential impact, but only whether the work is methodologically sound.

PLoS have decoupled two functions traditionally bundled together in the same journal. Specifically, they separate the assess-significance part of certification from the assess-soundness part. Methods and formal rigor are assessed conventionally. But the assess-significance component is done in a novel way, after publication, by tracking a variety of "*Article-level metrics*" including social bookmarking, blogging, and citation at the article level, then displaying this with the article. This innovative approach to part of assessment is only possible because PLoS One uncoupled

two certification sub-functions from one another, allowing the functions to be performed by different structures.

PLoS One also decouples the copyediting function; its *author guidelines page* warns that manuscripts "will not be subject to detailed copyediting. Obtaining this service is the responsibility of the author." But PLoS does not simply assume articles will be perfectly edited; instead, the guidelines give a list of 21 external services that perform this function for a fee. As in the case of certification for importance, PLoS treats copyediting as a module than can be decoupled and run independently.

This approach has been very successful for PLoS One; according to figures available on *their website*, they published over 5000 articles in 2010, and the rate at which new articles are published continues to grow. It has also been profitable, as authors (or their funders) pay a publication fee of US\$1350 per article. This success has not gone unnoticed by other publishers, who—despite early criticism (Butler, 2008)—have introduced similar "inclusive journals" (Wager, 2011) like *BMJ Open*, *Scientific Reports*, and *Sage Open*. However, this model still clings to some of the flaws in the traditional journal structure. First, publishing in PLoS One is exclusive; authors publish there, and only there. Neither do authors have choices about what kind of review they will receive. They may wonder if they could get better value for their money, as PLoS publishes an article for \$1350, while the ArXiv, which performs a much more limited editorial review, spends about \$7; as Poynder (2011) asks, "is the additional work undertaken by PLoS One 193 times more costly than ArXiv's moderation process?" Finally, one wonders whether a future scholarly journal ecosystem dominated by a few inclusive megajournals will not become as hidebound and oligarchic as the current system, dominated by a few publishers. Again we must wonder: what if we started here and then decoupled even more?

### **Post-publication review services**

A third scholarly communication structure that suggests the potential of the DcJ is the post-publication review service. There



are several of these in existence, but for the sake of space we will focus on two: Faculty of 1000 and Mathematics Reviews.

Faculty of 1000 (F1000), according *its website*:

... identifies and evaluates the most important articles in biology and medical research publications. The selection process comprises a peer-nominated global “Faculty” of the world’s leading scientists and clinicians who rate the best of the articles they read and explain their importance.

The goal of the service is to provide an additional filter, after classical peer review, to help researchers manage their growing reading lists. In doing so, they provide another example of a successful decoupled certification module. While some have argued that F1000 ranking correlate strongly with Thomson’s Journal Impact Factor (JIF) and are thereby of little value (Nature Neuroscience, 2005), Allen et al. (2009) show that F1000 does indeed spot valuable research overlooked by high-profile journals.

Mathematics Review is an abstracting service, but one that is occasionally called into service as a post-publication peer review venue when the traditional journals fail in their role as certifiers. In this case, abstracters may abandon objectivity and attack papers and their reviewers directly. As Kuperberg (2002) describes:

The community is often angry with the referees of [papers that should not have passed review], but anonymity protects them from the readers rather than the authors. Typically the Math Review sets the record straight.

In this way, Mathematics Review acts as certification’s second line of defense, a failsafe against the inevitable failures of the primary system.

These services and others like them are the most successful at decoupling the certification layer, because they do only that—unlike PLoS One or even the overlay journals, they make few if any attempts to perform other dissemination functions like marketing, search, or manuscript preparation. However, they cannot replace the current certification layer because they fail to sufficiently provide the indirect function of rewarding authors; again in Kuperberg’s (2002) words, “they are not designed to substitute for journal names in the author’s list of publications” (264). That is, they have decoupled part of the certification function, but not enough to fulfill it entirely.

### **The deconstructed journal**

This last example of decoupling is different from the other three because it has not actually been implemented as a working system. However, the Deconstructed Journal (Smith, 1999) is important to discuss because it remains the most complete description of the DcJ. Indeed, we see the DcJ as a way to implement Smith’s earlier vision, making a few modifications and taking advantage of advances in information technology over the last 12 years.

The Deconstructed Journal (DJ) is based on “three insights”:

1. We shouldn’t confuse the means (the journal) with the function.
2. Any replacement to journals must “satisfy the same needs” as current system.

3. This can be achieved by cooperating agencies; there’s no need for a central publisher.

The DJ decouples most of the functions of the journal, except those gathered in a “Subject Focal Point” (SFP), which brings together relevant literature and serves to as a portal for a community of readers. Archiving, preparation, and certification are all handled by specialist services. The SFP manages marketing and serves as a focal point for community-defining. This is a remarkably prescient vision, as it predates widespread adoption of many technologies that would greatly facilitate the DJ. Development of DOIs, OAI-PMH, IRs, social media, and other technologies makes this a significantly more practical and attractive framework, as Smith points out in a 2003 follow-up article.

van de Sompel et al. (2004) suggest many of the same ideas as Smith, using the ecosystem around the ArXiv subject repository as an example of a publishing value chain that is already partly “decomposed” (van de Sompel, 2000). They point out that a “loosely coupled” system has three major advantages: it encourages innovation, adapts well to changing scholarly practices, and democratizes the largely monopolized scholarly communication market.

However, the DJ and decomposed models do still have some weaknesses. Neither Smith’s nor van de Sompel’s proposals take into account the power of social media to convey scholarship. Today we can imagine collections of more diffuse social media communities, like the ones that form around Twitter hashtags, replacing Smith’s central SFPs. Second, and most importantly, neither Smith nor van de Sompel et al. spend much time laying out plans to gradually change from the present system to the ones they propose. This is entirely appropriate for these early proposals, which are quite revolutionary in scope. However, without next steps, the DJ will remain just a good idea.

### **THE FUNCTIONS OF THE DECOUPLED JOURNAL**

The DcJ (to distinguish it from Smith’s DJ), is an updating of Smith’s DJ, also incorporating similar suggestions from others including (Ginsparg, 2004; van de Sompel et al., 2004; Casati et al., 2007; Hendler, 2007; Cassella and Calvi, 2010). It takes full advantage of the Web’s growing power and pervasiveness to give authors and readers complete control over the scholarly objects they produce and consume, and gives service providers unprecedented freedom to specialize, mutate, and innovate.

The base unit of the DcJ is the scholarly object, which can be anything from a dataset or annotation to an article or monograph—anything scholars produce that they want to share. Instead of simply landing in one of thousands of vertical journal bins, this object ricochets around a rich ecosystem of modular services, acquiring new metadata, comments, stamps, links, citations, annotations, and edits as it goes. It is safely preserved and identified in long-term storage, mirrored all over the planet. It is indexed in general and specialized search engines and pushed to specialist readers eager for its specific content. It, and millions like it, forms a universal journal, but not one with any central publisher. This is a metajournal; like the web, it defines the smallest possible set of central structures and standards, then opens the

floodgates to the creativity and productiveness of thousands of service providers and millions of users.

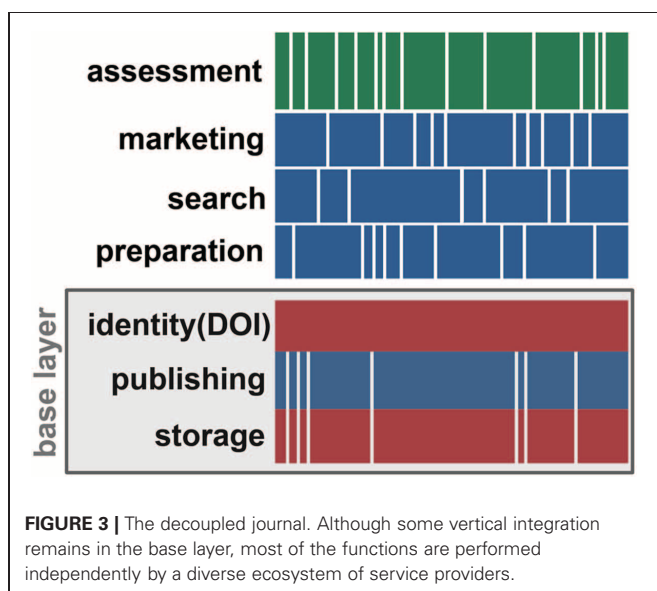
The best way to describe the DcJ, though, is to recall that in refactoring, we must make sure the system continues to perform all the functions it did before. So in this section, we will go through the functions of the journal one by one, describing its provision in the DcJ. First are the functions of archiving: persistent storage and persistent identification. Because these are all done at the same level, we will also discuss publication here. This is followed by a discussion of other journal functions, including preparation, search, marketing, and certification. We will replace this final term with “assessment,” reflecting that quality judgments in the DcJ will be subtler than simple binary yes/no stamps.

**Figures 3 and 4** describe the structure of the DcJ. In **Figure 3**, we see that the vertical silos have been replaced by horizontal bands of services, each performed by one or more independent service organizations. **Figure 4** gives an example of one way a given article might navigate this system.

### THE BASE LAYER: PERSISTENT STORAGE, PERSISTENT IDENTIFICATION, AND PUBLISHING

**Definition:** A permanent, open, web-accessible home for all scholarly products.

**Description:** This module is special, because as the base layer, all the others depend on it. In the DcJ, using a base layer service is the least possible action a scholar can take toward in sharing her work. The base layer is also special because it couples three functions into a single service. This is because refactoring is not about blindly decoupling every function in sight; rather, it is meant to reduce coupling as far as practical *but no further*. Long-term storage without persistent IDs means there is no sure way to find the item again: it's not storage, it's disposal. Similarly, there is no point in long-term identifiers if the identified object goes away. Finally, in this age of cheap and widespread connectivity, it is scarcely harder to store something online than off.



Moreover, making stored information objects networked is necessary for making mirrored backups at other sites, a crucial practice to safeguard data.

So the base layer publishes work, but we should not mistake this for “publishing” as the term is used today: reaching the end of a long submission, revision, and review process, then registering and disseminating an article in a journal. The DcJ turns that model on its head, making publication the first step in the process. It is a trivial step as easy as clicking a button, but one required to make further progress in meaningfully communicating a result of scholarly work.

**Who does it now:** Today, commercial and non-profit publishers handle storage and provision of a Document Object Identifier (DOI), a persistent identifier. Libraries may provide distributed backup storage in the form of paper copies, although this practice, at any volume, is certainly coming to an end. Publishers handle the electronic distribution of articles, and libraries (for now) distribute dead-tree copies. Growing number of articles are stored in freely accessible institutional and subject-area repositories.

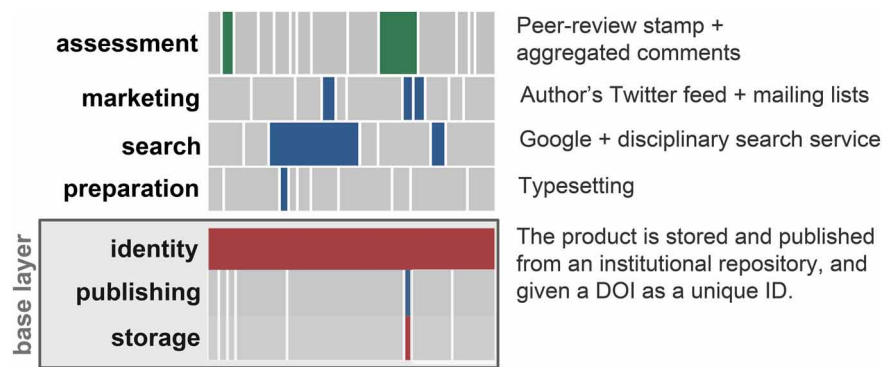
**Options in the future:** In the future, authors will be able to choose where to deposit their work. In most cases, they will likely choose free online repositories to store and publish their work, since these will be reliable, easy, and support the important other functions as well as their for-pay counterparts. They may even have an institutional mandate (Bosc and Harnad, 2005) to do so. However, if fee-based repository services can offer useful additional functions, these may emerge as well. For instance, a repository could support comprehensive versioning and “forking” of papers, making publishing more like open-source software development (Casati et al., 2007). Identification will probably continue, at least in the medium term, to be provided by the DOI system, which has shown itself to be scalable and effective. The biggest change in these functions is that authors will choose to deposit a larger variety of materials as upstream services evolve to add value to them. So, products like datasets, reviews, comments, notes, blog posts, and even tweets will all find their way into being persistently locatable and available on the Web.

**Transitional stages:** There is almost no transitional work needed for services of this kind; hundreds of institutional and subject-area repositories already exist, and continue to fill with articles. One change is that many of these do not mint DOIs, although they do provide relatively permanent identification with a URI. Another change is that authors will need convincing to deposit non-article items in repositories; this is already beginning to happen with products like datasets. Once non-article items can be peer reviewed by external modules, this process will diversify and accelerate.

### PREPARATION

**Definition:** Changing the format of a work to make it more suitable for a given (human or electronic) audience.

**Description:** There are many ways in which work needs to be transformed for dissemination. It may need copyediting, typesetting, or migration to alternate file formats. Datasets may need annotation or conversion to standard representations. Metadata may need to be added. Authors may want semantic markup to represent claims in machine-readable terms (Buckingham et al.,



**FIGURE 4 |** An example of a single article's view of the decoupled journal. Here we see one of many possible paths for an article in the decoupled journal. Authors and funders select which services and providers are best for a given article.

2000; Groth et al., 2010). In all these cases, the *content* of the work is unchanged, or nearly so; the representation is what is altered.

**Who does it now:** There are numerous companies that sell copyediting as a service to individual authors. However, preparation is still primarily the responsibility of the publishing journal, with authors expected to meet base guidelines. Although many journals outsource these tasks to specialists, saving money over doing conversions in-house, authors (or more often today, subscribers) do not get a say in whether that money is well-spent. What if, as an author, I want to pay to have my publication marked up entirely in RDF? Why should I pay for conversion to PDF if I think my readers only want HTML? Authors should be able to choose, based on their funding and desire, the forms their works will have.

**Options in the future:** In the DcL, authors will select the representations they prefer. An open market for these services, purchased à la carte, will drive down prices and reward the most valuable. Meanwhile, the open intellectual market will provide incentive for scholars to patronize preparation services whose work consistently broadens audiences and boost impact.

**Transitional stages:** PLoS One's policy of unloading copyediting to authors is a key precedent, and a step toward decoupling all preparation tasks from the other functions the metajournal. If more journals can be convinced to follow their lead, a market for preparation services of various types will continue to grow and diversify. As the cohesion and coupling of the scholarly communication ecosystem crumbles, this marketplace will be ready to accept the volume of papers and other products published by the metajournal.

## SEARCH

**Definition:** Connecting users to scholarly objects that meet their immediate information needs.

**Description:** To the best of our knowledge, search has not been suggested as one of the functions of the scholarly journal before. It is, however, increasingly indispensable. Over the last three decades scholars have been finding a growing percentage of their reading via search rather than browsing, a trend that still continues (Tenopir and King, 2008). Scholars have maxed out their ability to index work in their own heads, and so rely on the indexes

maintained by search engines as the size of the literature continues to grow.

**Who does it now:** Currently scholarly work is indexed and searched at journal, publisher, library, subject, and global levels. Some search services, like Elsevier's Scopus, or dozens or subject-specific indexes, are sold to subscribers. Others are available for free from libraries and repositories. Google Scholar is a free, ad-supported search service that has seen wide adoption. Our own experience and anecdotal evidence suggests that scholars are migrating toward more global search tools and away from those offered by journals or publishers.

**Options in the future:** The future of academic search is likely to look quite similar to today, with a variety of providers using different business models to search different bodies of literature. One change is that search engines will have to accommodate different types of scholarly products as these become more important. Another is that these search engines will begin to incorporate signals like downloads, comments, and links to make better relevance judgments. They will also incorporate information about a searcher's professional social networks to personalize results further. Finally, we will see search continue to supplant browsing; readers will replace pushed content to just-in-time information pulled from search engines.

**Transitional stages:** There will be little if any transitional stage between the future of scholarly search and its present; since search is already mostly decoupled from the other functions, it will freely evolve, driven by market forces.

## MARKETING

**Definition:** Distributing scholarly content to users who have an ongoing need for it.

**Description:** Marketing should not be thought of as merely a commercial enterprise. Publishers do market their journals to subscribers, but this is much less significant than scholars' use of journals to market their own ideas. In this regard, the *push* of marketing should be thought of as complementary to the *pull* of search. When the marketing function is working efficiently, scholarly products are seen by their maximum useful audience, and individual scholars regularly consume all and only the work that is most valuable to them.

**Who does it now:** Today journals are the pre-eminent marketing space for scholars' ideas. Most scholarly articles are useful only to an extremely limited audience; authors face the problem of marketing their work to this tiny group of potentially interested readers. Today's narrowly focused journals have evolved to be good at this. These journals benefit authors, but also readers, who have the complementary need to access to as much literature as possible in their narrow sub-specialties. The problem, though, is that no matter how thin the subject matter of a journal is sliced, there will always be papers on the edges of its coverage. Readers' interests will never perfectly match the content of a journal. Meanwhile, ever-finer divisions promote fractured, isolated, and disconnected research.

**Options in the future:** Curated subject-area hubs like Smith's (1999) SFPs will form narrowly focused, journal-like information markets that are entirely decoupled, serving to connect authors and readers but leaving certification, publication, and archiving to others. However, unlike in Smith's vision, in the DcJ these will take a backseat to an entirely different form of marketing feeds.

Feeds will be powered not by expert editorial decisions but by analysis of a user's professional social network and past preferences. They will use dozens of data sources to analyze the reading, bookmarking, downloading, commenting, and sharing of a scholarly community as well as the assessments assigned to articles and their sources comparing them to the same activities of a given user. Over time, this will allow the system to make intelligent recommendations, both for reading material and for colleagues to "follow." This has shown to be an effective way of creating strong but decentralized communities on services like Twitter.

Using informal ties to market and filter work is not a fundamentally new idea; scholarship has always been shaped by informal networks and "invisible colleges" (de Solla Price and Beaver, 1966). The true power of the scholarly social Web is not in formalizing or altering these ties (although this will happen); rather, it is in exposing them, uncovering the markers of "scientific 'street cred'" (Cronin, 2001) for use as inputs for a wide array of computational techniques. Google uses the humble web link to fuel algorithms that have made it the user interface for Web. Imagine using the aggregated information footprints from millions of scholars to make similarly useful recommendations on which research they will find useful and important. And of course, scholars will have choices of multiple recommendation systems, letting algorithms compete on coverage, efficiency, creativity, and price. Scholars will decide which feeds to consume the same way they decide what journals to read now: by seeing which ones consistently surface content that's valuable to them. Decoupling marketing from the journal's other functions allows the market to quickly assess and reward innovative, effective systems.

To market their work, authors will think less in terms of where to submit products, and more about building connections over social networks with those scholars who want to see the kind of work they do. Marketing services may spring up to meet this need, driven by people with unusually high degrees of connectivity across multiple communities. Their knowledge of the network will be available for a fee, their service resembling matchmaking more than traditional marketing.

### Transitional stages

The transition to a less centralized, more feed-based market has already begun. Tools like *Mendeley* and *CiteULike* already use network analysis of users' reading habits to tailor recommendations (Bogers and van den Bosch, 2008; Henning and Reichelt, 2008). Many scholars now turn to tools like Twitter feeds for marketing and being marketed to (Priem and Costello, 2010). One of this paper's authors has largely stopped reading journal tables of contents, finding that his Twitter feed gives him more relevant reading suggestions from a wider range of sources. In this environment, announcing a publication to one's feed is like publishing in a journal narrowly focused on the interests of your community. This will only become more pronounced as more scholars move more of their interactions online, and as data about these interactions accumulates.

### ASSESSMENT

**Definition:** Attaching an assessment of quality to a scholarly object.

**Description:** We use "assessment" instead of the more traditional term "certification" to reflect the broader, more nuanced evaluation performed in the DcJ. A great many approaches to this have been suggested. It is useful to organize all these approaches along a set of dimensions that are more or less orthogonal to one another. We suggest such a set below, containing four scalar and three binary dimensions:

- *Structure* is anchored on one side by free text with no structure, and on the other by the maximum structure, a yes/no dichotomy. Most peer reviews fall somewhere in between, although they ultimately resolve into a yes or no ruling. Online article commenting systems mostly produce unstructured text.
- *Anonymity* runs from complete anonymity to real names backed up by globally unique identifies like those proposed by the ORCID initiative.
- *Granularity* refers to the size of the unit being assessed, from individual words on unique versions to global comments on the whole of a single version.
- *Aggregation* can be at a level as small as a single review on each paper or large as tens of thousands of users' downloads, each representing a single yes/no assessment.
- *Invited or not:* Are reviews accepted from specific people, or anyone?
- *Assessing significance or not:* Do reviews assess soundness only, or do they make the more subjective judgment of significance?
- *Published or not:* Are reviews published, or kept secret?

We can imagine these oppositions as describing dimensions in n-dimensional space. Any type of review imaginable can in theory be represented by exactly one point in this space. Of course, we do not claim that these particular dimensions are the only way to break down the topic, but rather that some set of dimensions like this one is a useful way to describe the many forms of certification.

**Who does it now:** Today, most of the possible certification n-space is empty of living examples. With a few exceptions, certification huddles around one small point: reviews that are unpublished, assess significance, and are by invited reviewers. There are two or



three reviews that tend to examine both the paper as a whole and the quality of individual sections. Reviewers are anonymous, and reviews go from relatively unstructured free text at first, to a final ruling of thumbs up or down.

**Options in the future:** We argue that we do not need another grand scheme to revolutionize certification. Instead, we need a market where thousands of innovators, commercial and otherwise, can respond to the needs of authors and readers to evolve a new certification structure over time. For this to happen, certification must be entirely decoupled from the cost and distraction of supporting the other three functions, so that assessment services may compete fairly and evolve quickly.

When certification agents finally see themselves as *certifiers* rather than as *publishers*, we expect to see substantially greater diversity in the certification ecosystem. Certainly we will see overlay-type services that continue to supply journal-like peer review and branding while publishing only collections of links to approved papers. However, freed from the burden and crutch of publishing, assessment projects will quickly innovate further, looking for ways to differentiate themselves from competitors. Certification n-space will experience a land rush, quickly filling as innovators look to stake out claims on new-and-improved models.

Assessment services will experiment with a wide variety of review types including soundness-only reviews, high-volume reviews, editorial-only reviews, double-blind reviews, published reviews, reviews that assign grades rather than pass/fail, specialized supplementary reviews for statistics or ethics, non-exclusive reviews, pooled reviewers, and paid reviewers. What they will all have in common is a need to attract cash or attention in a crowded marketplace. Some will market their services to authors, others to readers—both of whom benefit from certification. A few may even charge reviewers for the chance to publicize their views. Many scholars will no doubt be interested in creating their own systems, funded by their institutions or granting agencies. Services will compete based on prestige, cost, turnaround time, and quality of feedback; most will fail to find enough users to be relevant (or solvent), but some will flourish. These will have proven their worth.

Along with these “traditional” stamping organizations we will see more qualitative review services that gather comments on an article from across the Web (as the *Disqus* system now does for blogs). We will see crowdsourced reviews and wikified articles. We will also see services that support purpose-built commenting or annotation systems layered atop existing article storage. As suggested by Hemminger (2009) and others, comments or annotations would be first-class scholarly products that would themselves be plugged into the base layer and could be disseminated, marketed, and reviewed like any other scholarly object.

We will also see more quantitative, data-driven reviews. These will draw their inspiration from data-hungry companies like Google. They will draw their raw material from the once-invisible traces of scholarly activities that are increasingly leaving tracks in the medium of the Web: tracks like downloads, bookmarks, comments, tweets, blog posts, and citations. All these aggregated *altmetrics* data, along with information about the social network generating them, will be a resource of unprecedented predictive

power. We see early evidence for effectiveness of these approaches: webometrics techniques have delivered data on authors’ and institution’s productivity based on web mentions (Thelwall and Harries, 2004; Thelwall, 2008). Brody et al. (2006) are able to predict citation from early downloads, and Yan and Gerstein (2011) find that PLoS article-level metrics data from social sources resemble traditional citation data.

Recent studies have used Twitter activity to predict things like movie box-office earnings (Asur and Huberman, 2010) and stock prices (Bollen et al., 2010) with uncanny accuracy. Eventually, algorithmic prediction of articles’ impact may be similarly accurate; we do not yet know. The proof is in the pudding: if aggregated quantitative assessments can consistently pick articles that user’s value, their recommendations will become increasingly prestigious—and valuable. We can imagine a future in which administrators and funders value a certain time-tested, quantitatively based certification the same way they would value publication by a top journal today. After all, such quantitative metrics would be the result of aggregating many expert discussions and opinions together, rather than just two from reviewers.

There are two objections to this approach that deserve particular mention here. The first is, “do we really expect scholars to pay for services for review, then turn around and do reviews for the same services, for free?” The easy answer is, of course we do—it’s what scholars already do now for journals. The more accurate answer, though, is that this is just the sort of problem that market-based, decoupled review will be good at solving. Any payment to reviewers will be passed along as a charge to authors buying reviews; if they find that the reviews are better for it, then it will happen. And of course reviewers might be rewarded in ways other than money; published reviews, for instance, could accumulate various types of electronic and traditional citations that directly benefit their writers. Certainly, reviewers that consistently identify important papers early will have opportunity to profit from their prognostication, either monetarily or socially. Finally, we do not know how much reviews have to cost, since they have never been subject to market forces in isolation. Even the best estimates involve guesswork, and vary between US\$100 and 400 (Donovan, 1998; Rowland, 2002; Ware and Mabe, 2009). Competition is likely to drive these numbers down. Depending on the market’s elasticity, money-saving measures like automatic reviewer selection (Rodriguez et al., 2006) may become common. Perhaps small groups of scholars will create their own free rankings, relying on social networks to gather and manage the review process. We cannot know until we uncouple certification and let it respond to market pressures on its own.

A second objection is what is to keep wealthy authors (or their funders) from buying stamps outright? After all, this system seems built primarily around the needs of authors; is to keep them from exploiting the system at the expense of readers, who need stamps they can trust? Of course it would be possible to set up a stamping agency that cheerfully passed out stamps to the highest bidder. But then, what exactly would that purchaser be bidding for? As Smith (1999) puts it: “(corrupt) organizations would soon disappear as evaluators would have nothing to sell but their reputation” (84).



Underwriters' Laboratories (UL), an organization that charges manufacturers for product safety certifications, is a good example. Certainly UL are as susceptible to kickbacks as peer review stampers would be. But they have been trusted for over a century because the public knows they have more to lose by being corrupts than by being honest. Who would pay for a certification, once UL had been caught selling them?

It's important to note, though, that assessment services in the DcJ would not need to be commercial. Non-profits or individuals with time and inclination could make their own assessment environments, crawlers, and algorithms, potentially drawing on more trust from their communities. If scholars can do a better job of delivering consistently useful assessment than commercial enterprises, the latter will gradually fade away. The important thing is that *everyone* be given the opportunity and raw data to build assessment services, without the vast additional infrastructure of publication, marketing, copyediting, and other functions.

**Transitional stages:** The transition to a decentralized certification marketplace is the most challenging part of the move to a DcJ. Overlay journals are a logical step, although seem to have attracted little enthusiasm outside the narrow open access community. Perhaps better technical tools for managing overlays will change this. Another possibility is to extend overlay journals into areas unserved by their traditional counterparts. One could imagine a journal designed to add peer review to blog posts or research technical reports. Instead of encouraging small communities to create overlays, large publishers might be interested. After all, while they have the most to lose in the DcJ, they also have a lot to gain: their brands continue to carry value whether they operate as overlays or not. A major publisher moving one of its large titles to an overlay model would signal agility and innovation to competitors, subscribers, and authors alike, and allow the publisher to focus on their core product: certification. A third approach would be for post-publication services like F1000 to market their service more aggressively as a stamp that should sit alongside journal publications on a CV—after all, it does represent a review by peers. The biggest and most practical step forward in the short-term is to plant the provocative idea in the heads of publisher, authors, readers, and funders that journals exist mostly to provide certification. What if we let them *just do that*?

## WORKFLOW IN THE DECOUPLED JOURNAL

So far, these ideas are relatively abstract. Let's look at three imaginary examples of what scholarly communication might look like with the DcJ. Of course, these are just possibilities; the DcJ is meant to evolve, and one of its strengths is that we cannot predict exactly what it will look like.

### AUTHOR: ANA

Ana, a biologist, is finishing a study on Florida lizards. As she finishes a rough draft of her paper, she navigates to her institutional repository and saves it. She has a free account with NeoNote, an overlay system that interfaces with her IR to provide an interface for her and her colleagues to annotate and comment on her draft. She has another account with an aggregation service that brings in external comments about her posted papers from twitter or blog posts. She blogs that the draft is up for comment and in

a week both services have accumulated some interesting suggestions, criticisms, and annotations, which she works into a revised paper. Based on a commenter's suggestion, she sends this new version to StatStamp, a statistics review service, since she's using some relatively obscure techniques. The service gives her some advice with leads to a few minor changes, and then she's awarded a StatStamp seal of approval. This is recorded in the article's IR metadata, and also on a list of links maintained by StatStamp. She is now happy with the state of the article, so she submits to the most prestigious stamping agency in her field, Lizard Reviews. After a few rounds of reviews and revisions, each of which is published with links to her article, Ana gets her stamp. She's a bit disappointed that it is the "B" stamp instead of the "A" she was hoping for, but it's still a coup. Meanwhile, an argument she had with the reviewers has been picked up by LizardTalk, a conversation aggregator in her field. Several of her colleagues join in, and she meets a researcher from a different field whose expertise in Florida's lizard habitats makes him a perfect collaborator for her next study.

### AUTHOR: BEATRICE

Beatrice is a chemist in the middle of a large study. She has finished data collection, so she has uploaded her dataset to her institution's repository. She also decides to upload a paper explaining her preliminary findings. She pays out of her grant to have the paper's language polished up a bit, since she doesn't have time to write more than a draft. She also pays to have her claims encoded in several scholarly ontologies and attached to the article's metadata, so that machines can crawl, read and understand her conclusions and their warrant. Next week she gets an automated email from ChemCrawler, a bot that crawls chemistry papers. ChemCrawler combined her data with that of a researcher who did a similar study and found that the combined data both clearly disproves one of her claims, and also supports several new ones. She integrates the new data and claims, then decides that, since work is moving quickly in the area, she should publish to a wider audience. Her field's most prestigious stamp takes a while to get, so she submits to the cross-field stamping agency QuikStamp instead. This agency automatically assigns reviewers based on keywords and the author's social network; it also pays fast reviewers a bonus in credits they can use at a consortium of stamping agencies. This means that Beatrice gets her stamp in just a week. QuikStamp certifies thousands of papers every week, so there's little chance that someone will run into it. However, a small fee submits the newly stamped article to a service that pushes it to other scholars who will be most likely to find it valuable, based on their public browsing, download, bookmarking, and commenting profiles.

### READER: CARL

Carl is a medical researcher who reads his articles feed twice a day. In the morning, over coffee, he sets his aggregator to "must read," which delivers articles that are stamped by the American Medical Association and articles that are being heavily read and recommended by his social networks. In the evening, he switches to "up and coming," articles that haven't been stamped yet, but that his aggregator predicts will be the focus of conversation

in the next few weeks, based on the activity of early adopters in his network, early downloads, and host of other metrics. Over time the algorithm adapts to Carl's preferences using his input and its own prediction record as feedback. Carl typically comments on both stamped and unstamped articles—wherever the conversation looks interesting. Carl enjoys his stature in his small disciplinary community, which converses online the way it used to at conferences—informally, but willing to make strong arguments backed by research. Carl notices that one of his earlier comments has spawned a long and interesting discussion overnight, accumulating good metrics. It will now be automatically added to his CV as evidence of his leadership in the community.

## ADVANTAGES OF THE DECOUPLED JOURNAL

The DcJ has three major advantages over other schemes to reform scholarly publishing.

**Can be achieved incrementally:** The most important advantage the DcJ has over alternatives is that it can evolve gradually from the current system. Indeed, as we have seen, it is already beginning to do so. Just as in biological evolution, immense changes can occur *if* each step along the way is viable in its own right. The DcJ ensures this by continuing to fulfill the essential functions of the journal at every step. It never replaces the essential currency of the traditionally peer reviewed paper—it just promotes a system that allows this currency to evolve, giving alternate certification approaches the space to convince conservative decision-makers of their value. The DcJ offers extant publishers a chance to evolve as well, shedding their legacy function as “publishers” (which they do with tremendous inefficiency compared to simple web repositories) and becoming lean, responsive certification providers. Whether they will overcome institutional inertia to succeed in this is an open question; however, if they seize the opportunity, today's publishers' experience and reputations offer them an early lead over startup stamping services. The important thing is the DcJ gives these major stakeholders in scholarly communications something to do besides dig in and fight for their survival. In these ways the DcJ is a model that can be reached via progressive change.

**The decoupled journal is a paradigm shift:** Although the DcJ is achievable by evolutionary means, its ultimate result is a complete revolution in the scholarly communication system. As Smith (1999) notes of his DJ proposal, the complete unbundling of the journal's function is nothing less than a Kuhnian paradigm shift in the way we communicate science. This is important because such a change is overdue; it is naïve to expect a paradigm built around seventeenth-century technology meet the needs of the Information Age. Attempts to patch pieces of the system in isolation without addressing its fundamental anachronisms will founder. This is what we expect from tightly coupled systems, where change in one function affects all the others.

The DcJ offers a legitimate and fundamentally different alternative to the present system, an alternative rooted in the technologies and ethos of the current age: openness, diversity, connectedness, customization, decentralization, the power of data. It promises a relatively bloodless revolution, in which some of the skills and experience of the current players can be

gradually repurposed—but a revolution nonetheless. Nothing less will suffice.

The DcJ is in many ways similar to another well-known decoupled system with modest underpinnings but revolutionary implications, the Web. Both define a set of roles, and responsibilities for each role. Both maintain an effective central registry of IDs. But both systems provide little regulation beyond these minimal requirements. The DcJ, like the Web, embraces a *laissez-faire* approach to regulation, preferring to give the market the maximum possible space to innovate. This techno-anarchism has been extraordinarily effective in the case of the Web, allowing it to evolve functions far beyond its creators' dreams. This is no surprise, given that a central advantage of decoupling is the ability to freely adapt, modify and even occasionally break individual components without wrecking the system as a whole. There is good reason to suspect that the successful decoupling of the journal would lead to explosive innovation reminiscent of the Web's. The Web was itself invented as a scholarly communication platform (Clarke, 2010); it's time for us to reclaim that legacy.

**The decoupled journal empowers innovators:** It is worth repeating that the DcJ is not a scheme for reforming peer review, but rather a meta-scheme for creating a market to let peer review—and the journal's other functions—evolve. We believe this is necessary for three reasons. First, there is already no shortage of innovative ideas for the reform of peer review, and the list will continue to grow without our help. Second, and more importantly, these isolated ideas, whatever their merit, will never implemented at large-scale without fundamental change to the entire scholarly communication system. The current tight coupling between the four functions makes it very difficult to change one function without changing the others as well. When all the journal's functions are made available as modular services, though, new certification schemes will be able to clearly articulate value propositions, accurately price services, and realistically assess effectiveness—in short, they can sell themselves. This is a *sine qua non* for convincing scholars to embrace change in so fundamental an institution. Finally, we suggest that no scheme, no matter how well-conceived, will anticipate all scholars' requirements and concerns. Experience shows that is often better to favor adaptability and responsiveness over comprehensiveness and cleverness. Common sense also suggests that over time, a market that attracts hundreds or thousands of hungry innovators will prove more creative than any single individual. In the four centuries since the Scientific Revolution, we have seen the power of a decentralized, open market for scientific ideas (Franck, 1999). Sadly, our communication tools do share this approach; economist Mark McCabe describes the state of publishing as a “true market failure” (Poynder, 2002). We can fix this, simply by making individual functions available as individual services.

## CONCLUSION

The journal is built around the delivery of ink and paper by horses and boats. Today, we have better ink and faster horses, but no fundamental change. This change, especially in an institution as conservative as the academy, is not easy and takes time. We should not expect a fully decoupled metajournal to emerge in the

next year or even decade. However, neither should we expect the current system, based as it is on the paradigm of the seventeenth century, to continue with only small, evolutionary changes. There will be a revolution in scholarly communication, as the fundamental potential of the Web compared with traditional models puts increasing torsion on our system. The revolution will not be in the functions of the journal system, which have proven themselves over centuries, but on the structures of the system we use to perform them.

We suggest that this revolution will result in a more diverse and decentralized metajournal. In this DcJ, authors will publish any sort of product they create. They will adapt their work's form and make it retrievable with the help of external service providers. They will market it over richly connected networks with the help

of specialists or without. They will certify it in dozens of ways, using hundreds or thousands of competing stamping and ranking agencies and algorithms. And all this data will be managed, organized, and curated by a set of relevance and ranking tools that will present customized views of the metajournal for scholars, practitioners, and administrators alike.

The most sensible early steps to achieving this vision are for publishers and interested academics to begin selling peer review as a service that can be a one-for-one replacement for journal peer review. If this can be successful, it will establish a precedent for peer review decoupled from the other functions, giving more services of more kinds a chance to enter the market. This in turn will lead to greater awareness of this approach's advantages, gradually encouraging the academy to adopt the DJ.

## REFERENCES

- Allen, L., Jones, C., Dolby, K., Lynn, D., and Walport, M. (2009). Looking for landmarks: the role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLoS One* 4:e5910. doi: 10.1371/journal.pone.0005910
- Asur, S., and Huberman, B. A. (2010). "Predicting the future with Social Media," 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. (Toronto, ON, Canada), 492–499.
- Bogers, T., and van den Bosch, A. (2008). "Recommending scientific articles using citeulike," in *Proceedings of the 2008 ACM Conference on Recommender Systems*, (New York, NY: ACM), 287–290.
- Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *J. Comput. Sci.* 2, 1–8.
- Bosc, H., and Harnad, S. (2005). In a paperless world a new role for academic libraries: providing open access. *Learn. Publ.* 18, 95–100.
- Brody, T., Harnad, S., and Carr, L. (2006). Earlier Web usage statistics as predictors of later citation impact. *J. Am. Soc. Inf. Sci. Technol.* 57, 1060–1072.
- Brown, J. (2010). An introduction to overlay journals. Repositories Support Project: UK. Retrieved from <http://discovery.ucl.ac.uk/19081>. (in press).
- Buckingham, S. S., Motta, E., and Domingue, J. (2000). ScholOnto: an ontology-based digital library server for research documents and discourse. *Int. J. Digit. Libr.* 3, 237–248.
- Butler, D. (2008). PLoS stays afloat with bulk publishing. *Nature* 454, 11.
- Casati, F., Giunchiglia, E., and Marchese, M. (2007). Publish and perish: why the current publication and review model is killing research and wasting your money. *Ubiquity* 3. doi: 10.1145/1226694.1226695
- Cassella, M., and Calvi, L. (2010). New journal models and publishing perspectives in the evolving digital environment. *IFLA J.* 36, 7–15.
- Clarke, M. (2010, January 4). Why hasn't scientific publishing been disrupted already? *The Scholarly Kitchen*. Retrieved January 30, 2010, from <http://scholarlykitchen.sspnet.org/2010/01/04/why-hasnt-scientific-publishing-been-disrupted-already/>
- Cronin, B. (2001). Bibliometrics and beyond: some thoughts on web-based citation analysis. *J. Inf. Sci.* 27, 1–7.
- Ding, Y., Jacob, E. K., Caverlee, J., Fried, M., and Zhang, Z. (2009). Profiling social networks: a social tagging perspective. *D-Lib Magazine*, 15, 1082–9873.
- Donovan, B. (1998). The truth about peer review. *Learn. Publ.* 11, 179–184.
- Franck, G. (1999). Scientific communication: a vanity fair? *Science* 286, 53–55.
- Ginsparg, P. (1997). Winners and losers in the global research village. *Ser. Libr.* 30, 83–95.
- Ginsparg, P. (2004). Can peer review be better focused? *Sci. Technol. Libr.* 22, 5–17.
- Gordon, R., and Poulin, B. J. (2008, October 15). There is but one journal: the scientific literature. Retrieved from <http://www.plosmedicine.org/annotation/listThread.action?inReplyTo=info%3Adoi/10.1371/annotation/b70a4689-cf09-4db6-a97b-8608b87e629e&root=info%3Adoi/10.1371/annotation/b70a4689-cf09-4db6-a97b-8608b87e629e>
- Gotzsche, P. C., Delamothé, T., Godlee, F., and Lundh, A. (2010). Adequacy of authors' replies to criticism raised in electronic letters to the editor: cohort study. *Br. Med. J.* 341, c3926.
- Greaves, S., Scott, J., Clarke, M., Miller, L., Hannay, T., Thomas, A., and Campbell, P. (2006). Nature's trial of open peer review. *Nature* 444, 971.
- Groth, P., Gibson, A., and Velterop, J. (2010). The anatomy of a nanopublication. *Inf. Serv. Use* 30, 51–56.
- Hemminger, B. (2009). NeoNote: suggestions for a global shared scholarly ecosystem. *D-Lib Magazine* 15. doi:10.1045/may2009-hemminger
- Hendler, J. (2007). Reinventing academic publishing, part 2. *IEEE Intell. Syst.* 22, 2–3.
- Henning, V., and Reichelt, J. (2008). "Mendeley – a last.fm for research?" in *IEEE Fourth International Conference on eScience* (Indianapolis, IN: IEEE) 327–328.
- Jefferson, T., Rudin, M., Brodney Folse, S., Davidoff, F. (2007). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Systematic Reviews* 2007, Issue 2. Art. No.: MR000016.
- Kuperberg, G. (2002). Scholarly mathematical communication at a crossroads. *Nieuw Archief voor Wiskunde* 5, 262–264.
- Moyle, M., and Lewis, A. (2008). *RIOJA (Repository Interface to Overlaid Journal Archives) project: final report*. Retrieved from <http://discovery.ucl.ac.uk/12562/>.
- Nature Neuroscience. (2005). Revolutionizing peer review? *Nat. Neurosci.* 8, 397.
- Peters, D. P., and Ceci, S. J. (1982). Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behav. Brain Sci.* 5, 187–195.
- Poynder, R. (2011, March 7). PLoS ONE, open access, and the future of scholarly publishing. *Open and Shut?* Retrieved March 17, 2011, from <http://poynder.blogspot.com/2011/03/plos-one-open-access-and-future-of.html>
- Poynder, R. (2002). A true market failure. *Inf. Today* 19. Retrieved from <http://www.infotoday.com/it/dec02/poynder.htm>
- Priem, J., and Costello, K. L. (2010). "How and why scholars cite on Twitter," in *Proceedings of the 73rd ASIS&T Annual Meeting*. Presented at the American Society for Information Science and Technology Annual Meeting, (Pittsburgh, PA).
- Reade, C. (1989). *Elements of Functional Programming*. Boston, MA: Addison-Wesley.
- Rodriguez, M. A., Bollen, J., and van de Sompel, H. (2006). The convergence of digital libraries and the peer-review process. *J. Inf. Sci.* 32, 149–159.
- Roosendaal, H., and Geurts, P. (1997). Forces and functions in scientific communication: an analysis of their interplay. Cooperative Research Information Systems in Physics, August 31–September 4 1997. (Oldenburg, Germany). Retrieved from <http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html>
- Rowland, F. (2002). The peer-review process. *Learn. Publ.* 15, 247–258.
- Schriger, D. L., Chehrizi, A. C., Merchant, R. M., and Altman, D. G. (2011). Use of the Internet by print medical journals in 2003 to 2009: a longitudinal observational study. *Ann. Emerg. Med.* 57, 153–160. e3.
- Smith, J. (1999). The deconstructed journal: a new model for academic publishing. *Learn. Publ.* 12, 79–91.
- Smith, J. W. T. (2003). "The deconstructed journal revisited: a review of developments," in *Proceedings. Presented at the ICCI/IFIP*

- Conference on Electronic Publishing-ElPub03: From information to knowledge*. (Minho, Portugal). 2–88.
- de Solla Price, D. J., and Beaver, D. (1966). Collaboration in an invisible college. *Am. Psychol.* 21, 1011.
- Stevens, W. P., Myers, G. J., and Constantine, L. L. (1974). Structured design. *IBM Syst. J.* 13, 115–139.
- Tenopir, C., and King, D. W. (2008). Electronic journals and changes in scholarly article seeking and reading patterns. *D-Lib Magazine* 14. doi:10.1045/november2008-tenopir
- Thelwall, M. (2008). Bibliometrics to webometrics. *J. Inf. Sci.* 34, 605–621.
- Thelwall, M., and Harries, G. (2004). Do the Web sites of higher rated scholars have significantly more online impact? *J. Am. Soc. Inf. Sci. Technol.* 55, 149–159.
- van de Sompel, H. (2000). “Closing keynote address,” in *Presented at the Coalition for Networked Information Meeting, San Antonio, Texas, USA*. Retrieved from <http://www.slideshare.net/hvdsomp/the-roof-is-on-fire>
- van de Sompel, H., Payette, S., Erickson, J., Lagoze, C., and Warner, S. (2004). Rethinking scholarly communication. *D-Lib Magazine* 10, 1082–9873.
- Wager, L. (2011, March 22). Journals that dare not speak their name. *BMJ Group Blogs*. Retrieved May 17, 2011, from <http://blogs.bmj.com/bmj/2011/03/22/liz-wager-journals-that-dare-not-speak-their-name/>
- Ware, M., and Mabe, M. (2009). *The STM Report: an Overview of Scientific and Scholarly Journal Publishing*. Oxford: STM: International Association of Scientific, Technical and Medical Publishers.
- Wenneras, C., and Wold, A. (2008). “Nepotism and sexism in peer-review,” in *Women, Science, and Technology: a reader in feminist science studies*, ed M. Wyer (New York, NY: Taylor & Francis) 46–52.
- Willinsky, J. (2003). Scholarly associations and the economic viability of open access publishing. *J. Digit. Inf.* 4. Available at: <http://journals.tdl.org/jodi/article/view/104>
- Yan, K.-K., and Gerstein, M. (2011). The spread of scientific information: insights from the Web usage statistics in PLoS article-level metrics. *PLoS One* 6:e19917. doi: 10.1371/journal.pone.00199
- Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 12 August 2011; paper pending published: 11 November 2011; accepted: 16 March 2012; published online: 05 April 2012.

Citation: Priem J and Hemminger BM (2012) Decoupling the scholarly journal. *Front. Comput. Neurosci.* 6:19. doi: 10.3389/fncom.2012.00019

Copyright © 2012 Priem and Hemminger. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits noncommercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.





# Learning from open source software projects to improve scientific review

Satrajit S. Ghosh<sup>1\*</sup>, Arno Klein<sup>2</sup>, Brian Avants<sup>3</sup> and K. Jarrod Millman<sup>4</sup>

<sup>1</sup> McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>2</sup> New York State Psychiatric Institute, Columbia University, New York, NY, USA

<sup>3</sup> Department of Radiology, PICSL, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

<sup>4</sup> Helen Wills Neuroscience Institute, University of California Berkeley, Berkeley, CA, USA

## Edited by:

Nikolaus Kriegeskorte, Medical Research Council Cognition and Brain Sciences Unit, UK

## Reviewed by:

Harel Z. Shouval, University of Texas Medical School at Houston, USA  
Nikolaus Kriegeskorte, Medical Research Council Cognition and Brain Sciences Unit, UK

## \*Correspondence:

Satrajit S. Ghosh, McGovern Institute for Brain Research, Massachusetts Institute of Technology, 43 Vassar St., 46-4033F MIT, Cambridge, MA 02139, USA.  
e-mail: satra@mit.edu

Peer-reviewed publications are the primary mechanism for sharing scientific results. The current peer-review process is, however, fraught with many problems that undermine the pace, validity, and credibility of science. We highlight five salient problems: (1) reviewers are expected to have comprehensive expertise; (2) reviewers do not have sufficient access to methods and materials to evaluate a study; (3) reviewers are neither identified nor acknowledged; (4) there is no measure of the quality of a review; and (5) reviews take a lot of time, and once submitted cannot evolve. We propose that these problems can be resolved by making the following changes to the review process. *Distributing reviews to many reviewers* would allow each reviewer to focus on portions of the article that reflect the reviewer's specialty or area of interest and place less of a burden on any one reviewer. *Providing reviewers materials and methods to perform comprehensive evaluation* would facilitate transparency, greater scrutiny, and replication of results. *Acknowledging reviewers* makes it possible to quantitatively assess reviewer contributions, which could be used to establish the impact of the reviewer in the scientific community. *Quantifying review quality* could help establish the importance of individual reviews and reviewers as well as the submitted article. Finally, we recommend *expediting post-publication reviews* and *allowing for the dialog to continue and flourish* in a dynamic and interactive manner. We argue that these solutions can be implemented by adapting existing features from open-source software management and social networking technologies. We propose a model of an open, interactive review system that quantifies the significance of articles, the quality of reviews, and the reputation of reviewers.

**Keywords:** distributed peer review, code review systems, open source software development, post-publication peer review, reputation assessment, review quality

## INTRODUCTION

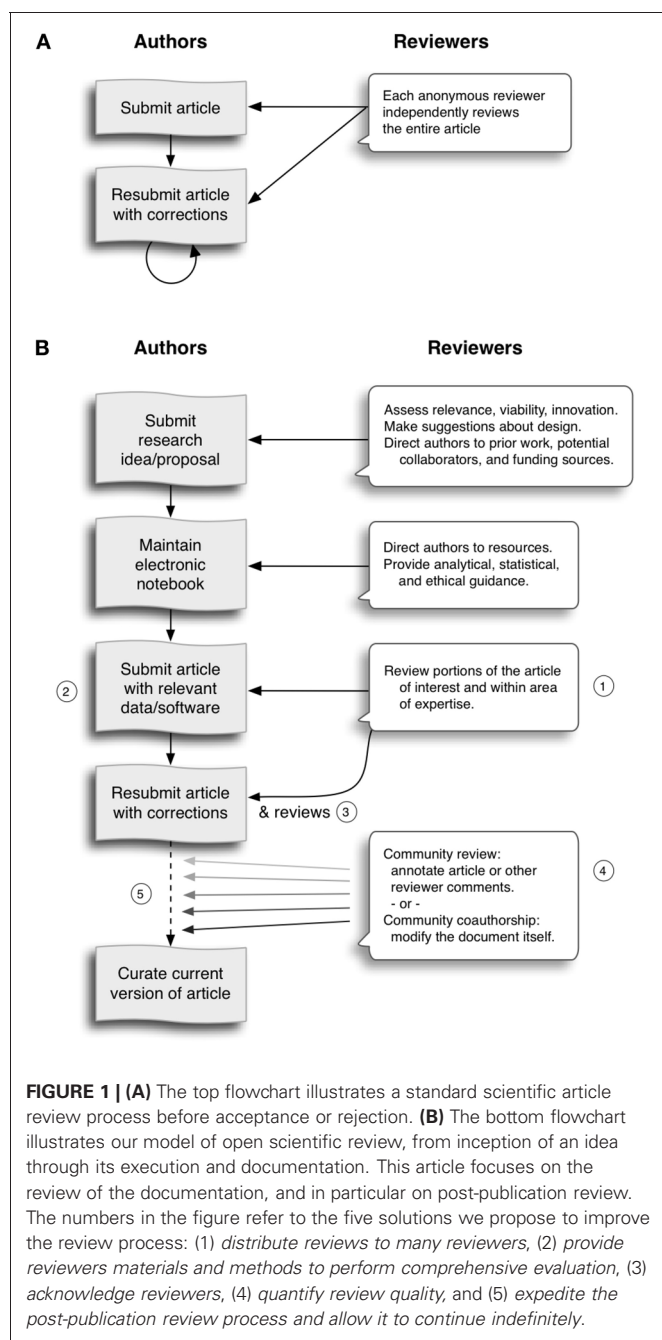
Scientific publications continue to be the primary mechanism for disseminating systematically gathered information about the natural world and for establishing precedence and credit for this research. In the current atmosphere of highly competitive and uncertain research funding, publications are instrumental in determining how resources are distributed, who gets promoted, and in which directions research advances. This has cultivated a publish-or-perish mentality where the focus is on maximizing the number of publications rather than on the validity and reproducibility of research findings, and a decrease in the amount of information apportioned to each article. Peer review is the primary means of filtering this rapidly growing literature prior to publication in an effort to ensure quality and validity.

Currently the typical review process for an article involves a preliminary screening by a journal editor followed by an anonymous and private review by a very small number of individuals (2–5, but often just 2) presumed to have expertise in the

research topic (**Figure 1A**)<sup>1</sup>. The editor takes into consideration the reviewers' recommendations to either publish, reject, or request revisions of the article. If published, the public only sees the final version of the article without any of the reviews (however, see, BioMed Central). After publication, problems such as fraud or mistakes are addressed via retraction after disclosure or exposure by countering articles or letters to the editor

<sup>1</sup>Currently, reviewers are solicited by the editors of journals based on either names recommended by the authors who submitted the article, the editors' knowledge of the domain or from an internal journal reviewer database. This selection process results in a very narrow and biased selection of reviewers. An alternative way to solicit reviewers is to broadcast an article to a larger pool of reviewers and to let reviewers choose articles and components of the article they want to review. These are ideas that have already been implemented in scientific publishing. The Frontiers system (frontiersin.org) solicits reviews from a select group of review editors and the Brain and Behavioral Sciences publication (<http://journals.cambridge.org/action/displayJournal?jid=BBS>) solicits commentary from the community.





(e.g., Chang et al., 2006; <http://retractionwatch.wordpress.com>). Through peer review and the scientific community's history of policing itself, scientists hope to achieve a self-correcting process. However, this self-correction is currently impeded by slow, private, and incremental reviews without objective standards and limited post-publication feedback. Without a transparent and objective framework, journals have gained a hierarchical stature, with some attracting the best authors, articles, and reviewers. These journals have been quantified by impact factors (Garfield, 1955), and as such, have overtaken the review process as arbiters of quality and significance of research. With the difficulty for individual reviewers to review the increasing number

and complexity of articles, and the use of journal impact factors as proxies for evaluations of individual articles, the integrity of the review process and, indeed, of science suffers (Smith, 2006; Poschl and Koop, 2008).

In contrast to peer review of scientific articles, when software programmers develop open source software and review their code, the process is open, collaborative, and interactive, and engages many participants with varying levels of expertise. There is a clear process by which comments get addressed and new code gets integrated into the main project. Since computer programs are much more structured and objective than prose, it is more amenable to standardization and, therefore, to review. These code review systems also take advantage of some of the latest technologies that have the potential to be used for publication review. Despite all of these differences, the purpose of code review systems mirror the purpose of publication review to increase the clarity, reproducibility, and correctness of contributions.

The most prominent example of a post-publication review system, arXiv.org, comes from the field of high energy particle physics. It has transformed the way results are disseminated, reviewed, and debated. Authors submit articles to arXiv even before they are submitted or appear in a traditional journal. Often, discussion and responses take place before the article appears in print. Interesting findings and the scientific discourse related to these findings are thus brought to the immediate attention of the community and the public. This process of rapid, fully open debate based on the exchange of technical preprints takes place even for major new results that in other fields would typically be shrouded in secrecy. A recent example was the open discussion of the possible discovery of a new particle at Fermilab's Tevatron accelerator that did not fit the Standard Model of particle physics<sup>2</sup>. However, this system has been applied to narrow domains of expertise, does not have a rating mechanism and its scalability in the context of increasingly interdisciplinary domains remains untested.

The advent of social networking technology has altered the traditional mechanisms of discourse, but the ease of adding to online discussions has also resulted in increasingly redundant and voluminous information. Blogs (e.g., polymathprojects.org), social network sites (e.g., Facebook, Google+) and scientific discussion forums (e.g., metaoptimize.com, mathoverflow.net, and researchgate.net) are redefining the technologies that extract, organize, and prioritize relevant, interesting and constructive information and criticism. In the scientific world, new discoveries and technologies make rapid dissemination and continued reappraisal of research an imperative. However, the scientific establishment has been slow to adopt these social technologies. The peer review system is one area where the scientific community may benefit from adopting such technologies.

For the publication review process to continue to play a critical role in science, there are a number of problems that need to be addressed. In this article, we list five problems and potential solutions that derive from distributed code review in open source software development.

<sup>2</sup><http://arstechnica.com/science/news/2011/05/evidence-for-a-new-particle-gets-stronger-ars>

## PROBLEMS WITH THE CURRENT PEER-REVIEW PROCESS

### REVIEWERS ARE EXPECTED TO HAVE COMPREHENSIVE EXPERTISE

Reviewers are expected to work in isolation, unable to discuss the content of an article with the authors or other reviewers. When faced with an article that may be authored by half a dozen or more experts in their respective disciplines, how could a few reviewers be expected to have the range of expertise necessary to adequately understand and gauge the significance (or insignificance) of all aspects of a given article? Why are the different components of an article, including the background, experimental design, methods, analysis of results, and interpretations handed over as a package to each reviewer, rather than delegated to many experts in each domain? Realistically, it is common practice for a reviewer to criticize portions of an article that he or she understands, is interested in, has time to read, and takes issue with, while falling silent on the rest of the article. This leads an editor to assume these silences are indicators of tacit approval. The unrealistic expectations placed on each of the reviewers, coupled with the delayed and sequential interactions they have with the authors and editors, have made the review process inefficient.

### REVIEWERS DO NOT HAVE SUFFICIENT ACCESS TO METHODS AND MATERIALS TO EVALUATE A STUDY

The typical review process does not require submission of data or software associated with an article (Association for Computing Machinery Transactions on Mathematical Software was an early exception), and the descriptions provided in methods sections are often inadequate for replication. This makes it impossible for a reviewer, if so inclined, to fully evaluate an article's methods, data quality, or software, let alone to replicate the results of the study. Failing to expose the methods, data, and software underlying a study can lead to needless misdirection and inefficiency, and even loss of scientific credibility (Ioannidis, 2005). One example is the case of Geoffrey Chang, whose rigorous and correct experimental work was later retracted due to a software bug that undermined the paper's conclusions (Chang et al., 2006).

### REVIEWERS ARE NEITHER IDENTIFIED NOR ACKNOWLEDGED

Review is currently considered one's unpaid "duty" to maintain the standards and credibility of scientific research. There is little motivation for potential reviewers to participate in the review process; some motivation comes from the knowledge gained from as yet unpublished results. However, the current system does not acknowledge their services in a manner that could factor into their evaluations for promotion and funding opportunities. In addition to acknowledging a reviewer's contributions for the benefit of the reviewer, identifying a reviewer has many benefits to science and scientific discourse, including transparency of the review process and proper attribution of ideas.

### THERE IS NO MEASURE OF THE QUALITY OF A REVIEW

Currently there is no way to objectively quantify the quality, strength, impartiality, or expertise of the reviews or reviewers. Without measures associated with the quality of any portion of a review, the community is forced to trust the qualitative assessment of the editor and the journal's impact factor as proxies for

quality. This prevents external scrutiny and makes it impossible to evaluate or standardize the review process.

### REVIEWS TAKE A LOT OF TIME AND ONCE SUBMITTED CANNOT EVOLVE

A lengthy review process holds up grant submissions, funding of research programs, and the progress of science itself. And even after this process, for the vast majority of articles none of the information (criticism or feedback) generated during the review is made publicly available (BioMed Central is one counterexample). Furthermore, after an article has been published, the review process simply ends even for those who participated, as if the work and interpretations of the results are sealed in a time capsule. Data, methods, analysis, and interpretations of the results are all a product of their time and context, and at a later time may not stand up to scrutiny or may yield new insights.

## PROPOSED RE-DESIGN OF THE PEER REVIEW PROCESS

There are notable examples of journals (e.g., Frontiers—frontiersin.org, BioMedCentral—biomedcentral.com, PLoS One—plosone.org) that address one or another of the above problems, but the vast majority of journals do not address any of the above problems. We propose an open post-publication review system for scientific publishing that draws on the ideas, experience, and technologies recently developed to support community code review in open source software projects.

**Figure 1B** illustrates this model of open scientific review, from inception of an idea through its execution and documentation. The numbers in the figure refer to the five solutions we propose to improve the review process that addresses each of the problems listed in the prior section: (1) distribute reviews to many reviewers, (2) provide reviewers materials and methods to perform comprehensive evaluation, (3) acknowledge reviewers, (4) quantify review quality, and (5) expedite the post-publication review process and allow it to continue indefinitely. With the continued inclusion of new comments, the concept of a "publication" itself gives way to a forum or an evolving dialogue. In this sense, review can be seen as a form of co-authorship. The end-to-end review process in **Figure 1B** would integrate collaborative authoring and editing (e.g., Google docs; annotum.org—Leubsdorf, 2011), reviewing and discussion of scientific ideas and investigations. This article focuses on the review of the documentation, and in particular on post-publication review.

In this section, we describe our proposed solutions, then highlight the relevance of current code review systems in addressing the problem and finally describe enhancements to the current systems to support our proposed solution.

### DISTRIBUTE REVIEWS TO MANY REVIEWERS

Reviewers would no longer work in isolation or necessarily in anonymity, benefiting from direct, dynamic, and interactive communication with the authors and the world of potential reviewers. This would help reviewers to clarify points, resolve ambiguities, receive open collegial advice, attract feedback from people well outside of the authors' disciplines, and situate the discussion in the larger scientific community. Reviewers could also focus on portions of the article that reflect their expertise and interests;

but they would, of course, have the opportunity to provide feedback on an entire article. Furthermore, they would not be held responsible for every aspect of the article, leaving portions that they are not qualified or interested in for others and their silence would not be mistaken for tacit approval. This will lessen burden placed on any one reviewer, enabling a more comprehensive, timely and scientifically rigorous review. This would also expose which portions of an article were not reviewed.

In case there is a fear of disclosure prior to publication<sup>3</sup>, of an overwhelming amount of participation in a review where anyone could be a reviewer, or of a lack of consensus across reviewers, there are at least three types of alternatives available. One would be to assign certain reviewers as moderators for different components of the article, to lessen the burden on the editor. A second would be to restrict the number of reviewers to those solicited from a pool of experts. This would still improve scientific rigor while lessening the burden on each individual reviewer, as long as they review specific components of the article they are knowledgeable about. A third would be to conduct a preliminary review consisting of a limited, possibly anonymous and expedited review process *prior to the full and open review* as we propose. At different stages of such a tiered review, reviewers might be assigned different roles, such as mediator, editor, or commenter.

### Relevance of code review systems

In the same manner that articles are submitted for review and publication in journals, code in collaborative software projects is submitted for review and integration into a codebase. In both scientific research and in complex software projects, specialists focus on specific components of the problem. However, unlike scientific review, code review is not limited to specialists. When multiple pairs of eyes look at code, the code improves, bugs are caught, and all participants are encouraged to write better code. Existing code review systems such as Gerrit (<http://code.google.com/p/gerrit>) as well as the collaborative development and code review functionality provided by hosting services like GitHub (<http://github.com>) are built for a distributed review process and provide reviewers the ability to interact, modify, annotate and discuss the contents of submitted code changes.

Indeed, the purpose of these systems mirror the purpose of scientific review—to increase the clarity, reproducibility and correctness of works that enter the canon. While no journals provide a platform for performing such open and distributed review, the Frontiers journals do provide an interactive, but non-public discussion forum for authors and reviewers to improve the quality of a submission after an initial closed review. In GitHub, code is available for everyone to view and for registered GitHub members to comment on and report issues on through an interactive web interface. The interface combines a discussion forum that allows inserting comments on any given line of code together with a mechanism for accepting new updates to the code that fix unresolved issues or address reviewer comments (an example is shown in Appendix **Figure A1**). These interactive

discussions become part of a permanent and open log of the project.

### Enhancing code review systems for article review

These existing code review systems, while suitable for code, have certain drawbacks for reviewing scientific articles. For example, the GitHub interface allows line-by-line commenting which reflects the structure of code. But commenting on an article's text should follow the loose structure of prose with comments referring to multiple words, phrases, sentences or paragraphs rather than whole lines. These comments should also be able to refer to different parts of an article. For example, a reviewer might come across a sentence in the discussion section of an article that contradicts two sentences in different parts of the results section. The interface should allow reviewers to expose contradictions, unsubstantiated assumptions, and other inconsistencies across the body of an article or across others' comments on the article. This system can be used in both a traditional review-and-revise model as well as a collaborative Wikipedia-style revision model that allows collaborative revision of the article. Since metrics keep track of both quality and quantity of contributions (discussed later), such an approach encourages revisions to an article that improve its scientific validity instead of a new article. A mock-up of such a review system is shown in **Figure 2**.

### PROVIDE REVIEWERS MATERIALS AND METHODS TO PERFORM COMPREHENSIVE EVALUATION

In a wide-scale, open review, descriptions of experimental designs and methods would come under greater scrutiny by people from different fields using different nomenclature, leading to greater clarity and cross-fertilization of ideas. Software and data quality would also come under greater scrutiny by people interested in their use for unexpected applications, pressuring authors to make them available for review as well, and potentially leading to collaborations, which would not be possible in a closed review process.

We propose that data and software (including scripts containing parameters) be submitted together with the article. This not only facilitates transparency for all readers including reviewers but also facilitates reproducibility and encourages method reuse. Furthermore, several journals (e.g., Science—[sciencemag.org](http://sciencemag.org), Proceedings of the National Academy of Sciences—[pnas.org](http://pnas.org)) are now mandating availability of all components necessary to reproduce the results (Drummond, 2009) of a study as part of article submission. The journal Biostatistics marks papers as providing code [C], data [D], or both [R] (Peng, 2009).

While rerunning an entire study's analysis might not currently be feasible as part of a review, simply exposing code can often help reviewers follow what was done and provides the possibility to reproduce the results in the future. In the long run, virtual machines or servers may indeed allow standardization of analysis environments and replication of analyses for every publication. Furthermore, including data with an article enables readers and reviewers to not only evaluate the quality and relevance of the data used by the authors of a study, but also to determine if the results generalize to other data. Providing the data necessary to reproduce the findings allows reviewers to

<sup>3</sup>To allay concerns over worldwide pre-publication exposure, precedence could be documented by submission and revision timestamps acknowledging who performed the research.





**FIGURE 2 | This schematic illustrates color-coded ratings assigned to text in an article or reviewer comment.** Such a visualization could help authors, reviewers, and editors quickly assess how much of and how favorably an article has been reviewed, and could be useful in a publishing model where an article is considered published after it garners a minimum rating over an appreciable amount of its content. **(A)** A reviewer selects some text which launches a colorbar for rating the text and a comment box, and **(B)** gives a low rating (red) for the text and adds a negative comment (a thumbs down appears in the comment box to reflect the rating). **(C)** Another reviewer selects the same block of text (which launches a comment box), then rates the text and some of the other comments. A red or blue background

indicates a cumulative negative or positive rating. In this example, the positive ratings outweigh that of the initial negative comment, turning the text from red to blue. Each reviewer's vote can be weighted by the ratings received by that reviewer's past contributions to peer review. **(D)** A reviewer selects the bottom-most comment to leave a comment about it. **(E)** The middle row shows how the ratings of an article's text can change over time. **(F)** The bottom row represents a dashboard summary of the ratings assigned to an article, including reviewer activity, coverage, and variation of opinion regarding the article. General comments can also be added for the article as a whole via the dashboard. The dashboard also indicates whether code, data and/or a virtual machine are available for reproducing the results of the article.

potentially drill down through the analysis steps—for example, to look at data from each preprocessing stage of an image analysis pipeline.

### Relevance of code review systems

While certain journals (e.g., PLoS One, Insight Journal) require code to be submitted for any article describing software or

algorithm development, most journals do not require submission of relevant software or data. Currently, it is considered adequate for article reviewers to simply read a submitted article. However, code reviewers must not only be able to read the code, they must also see the output of running the code. To do this they require access to relevant data or to automated testing results. Code review systems are not meant to store data, but

complement such information by storing the complete history of the code through software version control systems such as Git (git-scm.com) and Mercurial (mercurial.selenic.com). In addition to providing access to this history, these systems also provide other pertinent details such as problems, their status (whether fixed or not), timestamps and other enhancements. Furthermore, during software development, specific versions of the software or particular files are tagged to reflect milestones during development. Automated testing results and detailed project histories provide contextual information to assist reviewers when asked to comment on submitted code.

### Enhancing code review systems for article review

As stated earlier, code review systems are built for code, not for data. Code review systems should be coupled with data storage systems to enable querying and accessing code and data relevant to the review.

### ACKNOWLEDGE REVIEWERS.

When reviewers are given the opportunity to provide feedback regarding just the areas they are interested in, the review process becomes much more enjoyable. But there are additional factors afforded by opening the review process that will motivate reviewer participation. First, the review process becomes the dialogue of science, and anyone who engages in that dialogue gets heard. Second, it transforms the review process from one of secrecy to one of engaging social discourse. Third, an open review process makes it possible to quantitatively assess reviewer contributions, which could lead to assessments for promotions and grants. To acknowledge reviewers, their names (e.g., Frontiers) and contributions (e.g., BioMed Central) can be immediately associated with a publication, and measures of review quality can eventually become associated with the reviewer based on community feedback on the reviews.

### Relevance of code review systems

In software development, registered reviewers are acknowledged implicitly by having their names associated with comments related to a code review. Systems like Geritt and GitHub explicitly list the reviewers participating in the review process. An example from Geritt is shown in supplementary **Figure A2**.

In addition, certain social coding websites (e.g., ohloh.net) analyze contributions of developers to various projects and assign

“kudos” to indicate the involvement of developers. **Figure 3** shows an example of quantifying contributions over time. Neither of these measures necessarily reflect the quality of the contributions, however.

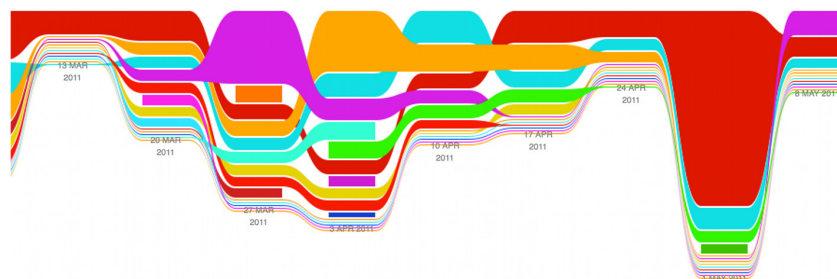
### Enhancing code review systems for article review

The criterion for accepting code is based on the functionality of the final code rather than the quality of reviews. As such, code review systems typically do not have a mechanism to rate reviewer contributions. We propose that code review systems adapted for article review include quantitative assessment of the quality of contributions of reviewers. This would include a weighted combination of the number (**Figure 3**), frequency (**Figure 4**), and peer ratings (**Figure 2**) of reviewer contributions. Reviewers need not be the only ones to have an impact on other reviewers’ standing. The authors themselves could evaluate the reviewers by assigning impact ratings to the reviews or segments of the reviews. These ratings can be entered into a reviewer database, referenced in the future by editors and used to assess contributions to peer review in the context of academic promotion. We acknowledge some reviewers might be discouraged by this idea, thus it may be optional to participate.

### QUANTIFY REVIEW QUALITY

Although certain journals hold a limited discussion before a paper is accepted, it is still behind closed doors and limited to the editor, the authors, and a small set of reviewers. An open and recorded review ensures that the role and importance of reviewers and information generated during the review would be shared and acknowledged. The quantity and quality of this information can be used to quantitatively assess the importance of a submitted article. Such quantification could lead to an objective standardization of review.

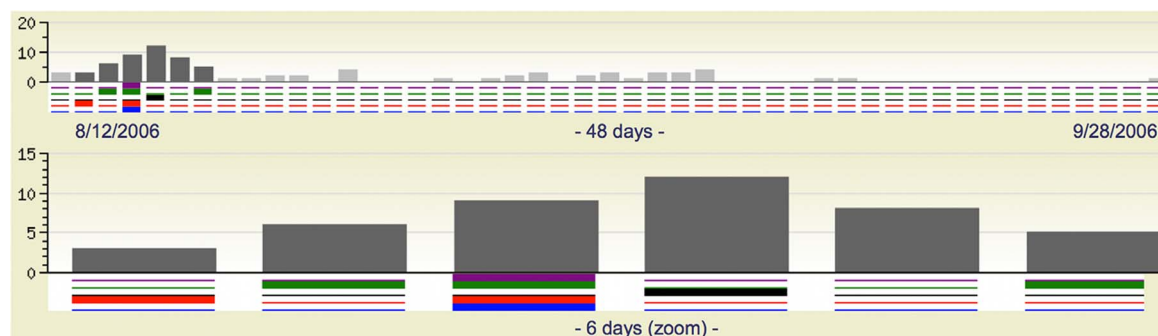
There exist metrics for quantifying the importance of an author, article, or journal (Hirsch, 2005; Bollen et al., 2009), but we know of no metric used in either article review or in code review for quantifying the quality, impact, or importance of a review, of a comment on a review, or of any portions thereof. Metrics have many uses in this context, including constructing a dynamic assessment of individuals or ideas for use in promotion and allocation of funds and resources. Metrics also make it possible to mine reviews and comment histories to study the process of scientific publication.



**FIGURE 3 | Example of a metric for quantifying contributions over time.** This is a screenshot of a ribbon chart visualization in GitHub of the history of code additions to a project, where each color

indicates an individual contributor and the width of a colored ribbon represents that individual’s “impact” or contributions during a week-long period.





**FIGURE 4 | Example of a metric for quantifying contributor frequency.** Quotes over Time ([www.qovert.info](http://www.qovert.info)) tracked the top-quoted people from Reuters Alertnet News on a range of topics, and

presents their quotes on a timeline, where color denotes the identity of a speaker and bar height the number of times the speaker was quoted on a given day.

### Relevance of code review systems

In general, code review systems use a discussion mechanism, where a code change is moderated through an iterative process. In the context of code review, there is often an objective criterion—the code performs as expected and is written using proper style and documentation. Once these standards are met, the code is accepted into the main project. The discussion mechanism facilitates this process. Current code review systems do not include quantitative assessment of the quality of reviews or the contributions of reviewers.

### Enhancing code review systems for article review

The classic “Like” tally used to indicate appreciation of a contribution in Digg, Facebook, etc., is the most obvious measure assigned by a community, but it is simplistic and vague. In addition to slow and direct measures of impact such as the number of times an article is cited, there are faster, indirect behavioral measures of interest as a proxy for impact that can be derived from clickstream data, web usage, and number of article downloads, but these measures indicate the popularity but not necessarily quality of articles or reviews.

We propose a review system (Figure 2) with a “reputation” assessment mechanism similar to the one used in discussion forums such as [stackoverflow.net](http://stackoverflow.net) or [mathoverflow.net](http://mathoverflow.net) in order to quantify the quality of reviews. These sites provide a web interface for soliciting responses to questions on topics related to either computer programming or mathematics, respectively (supplementary Figure A3). The web interface allows registered members to post or respond to a question, to comment on a response, and to vote on the quality or importance of a question, of a response, or of a comment. In our proposed review system, such a vote tally would be associated with identified, registered reviewers, and would be only one of several measures of the quality of reviews (and reviews of reviews) and reviewers. Reviews can be ranked by importance (weighted average of ratings), opinion difference (variance of ratings) or interest (number of ratings). Reviewer “reputation” could be computed from the ratings assigned by peers to their articles and reviews.

It would also be possible to aggregate the measures above to assess the impact or importance of, for example, collaborators,

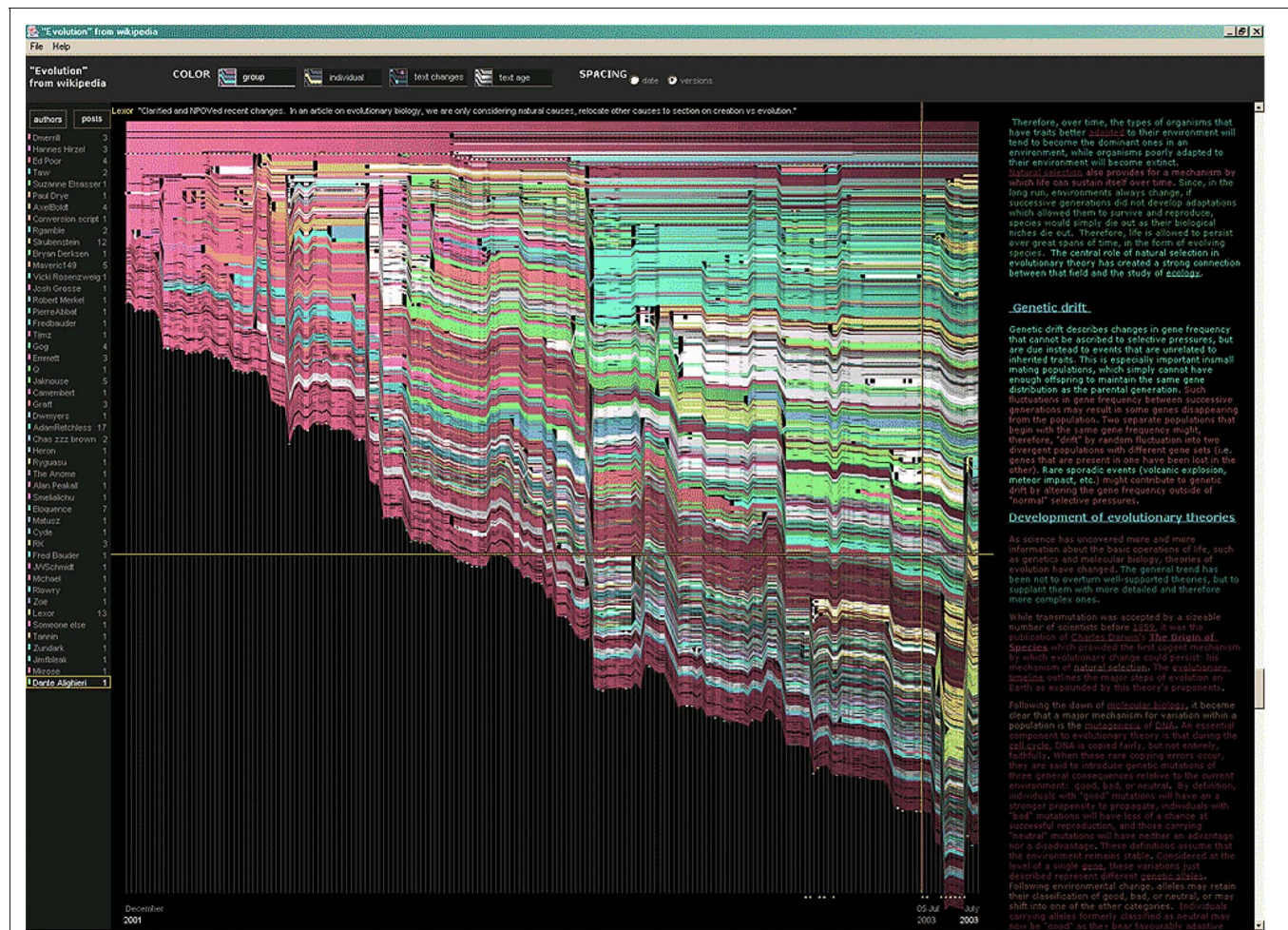
coauthors, institutions, or different areas of multidisciplinary research. As simple examples, one could add the number of contributions by two or more coders in Figure 3 or the number of quotations by two or more individuals in Figure 4. This could be useful in evaluating a statement in an article in the following scenario. Half of a pool of reviewers A agrees with the statement and the other half B disagrees with the statement. Deciding in favor of group A would be reasonable if the aggregate metric evaluating A’s expertise on the statement’s topic is higher than that of B. However, such decisions will only be possible once this system has acquired a sufficient amount of data about group A and B’s expertise on reviewing this topic, where expertise is related to the “reputation” assessment mentioned above.

### EXPEDITE REVIEWS AND ALLOW FOR CONTINUED REVIEW.

Once open and online, reviews can be dynamic, interactive, and conducted in real time (e.g., [Frontiers](http://Frontiers)). And with the participation of many reviewers, they can choose to review only those articles and components of those articles that match their expertise and interests. Not only would these two changes make the review process more enjoyable, but they would expedite the review process. And there is no reason for a review process to end. Under post-publication review, the article can continue as a living document, where the dialogue can evolve and flourish (see Figure 5), and references to different articles could be supplemented with references to the comments about these articles, perhaps as Digital Object Identifiers (<http://www.doi.org>), firmly establishing these communications within the dialogue and provenance of science, where science serves not just as a method or philosophy, but as a social endeavor. This could make scientific review and science a more welcoming community.

### Relevance of code review systems

Code review requires participation from people with differing degrees of expertise and knowledge of the project. This leads to higher quality of the code as well as faster development than individual programmers could normally contribute. These contributions can also be made well beyond the initial code review allowing for bugs to be detected and improvements to be made by new contributors.



**FIGURE 5 | A visualization of the edit history of the interactions of multiple authors of a Wikipedia entry ("Evolution").** The text is in the right column and the ribbon chart in the center represents the text edits over

time, where each color indicates an individual contributor ([http://www.research.ibm.com/visual/projects/history\\_flow/gallery.htm](http://www.research.ibm.com/visual/projects/history_flow/gallery.htm), Viegas et al., 2004).

### Enhancing code review systems for article review

Current code review systems have components for expedited and continued review. Where they could stand to be improved is in their visual interfaces, to make them more intuitive for a non-programmer to quickly navigate (Figure 2), and to enable a temporal view of the evolutionary history of an arbitrary section of text, analogous to Figure 5 (except as an interactive tool). As illustrated in Figure 1B and mentioned in the Discussion section below, co-authorship and review can exist along a continuum, where reviewers could themselves edit authors' text in the style of a wiki (e.g., [www.wikipedia.org](http://www.wikipedia.org)) and the authors could act as curators of their work (as in [www.scholarpedia.org](http://www.scholarpedia.org)).

## DISCUSSION

The current review process is extremely complex, reflecting the demands of academia and its social context. When one reviews a paper, there are considerations of content, relevance, presentation, validity, as well as readership. Our vision of the future of scientific review aims to adopt practices well-known in other

fields to reliably improve the review process, and to reduce bias, improve the quality, openness and completeness of scientific communications, as well as increase the reproducibility and robustness of results. Specifically, we see hope in the model of review and communication used by open source software developers, which is open, collaborative, and interactive, engaging many participants with varying levels of expertise.

In this article, we raised five critical problems with the current process for reviewing scientific articles: (1) reviewers are expected to have comprehensive expertise; (2) reviewers do not have sufficient access to methods and materials to evaluate a study; (3) reviewers are neither identified nor acknowledged; (4) there is no measure of the quality of a review; and (5) reviews take a lot of time, and once submitted cannot evolve. We argue that we can address all of these problems via an open post-publication review process that engages many reviewers, provides them with the data and software used in an article, and acknowledges and quantifies the quality of their contributions. In this article, we described this process (Figure 1B) together with a quantitative commenting

mechanism (**Figure 2**). We anticipate that such a system will speed up the review process significantly through simultaneous, distributed, and interactive review, an intuitive interface for commenting and visual feedback about the quality and coverage of the reviews of an article. The proposed framework enables measurement of the significance of an article, the quality of reviews and the reputation of a reviewer. Furthermore, since this system captures the entire history of review activity, one can refer to or cite any stage of this evolving article for the purpose of capturing the ideas and concepts embodied at that stage or quantifying their significance over time.

Despite the advantages of our proposed open review process and the promise offered by existing solutions in other domains, adopting the process will require a change of culture that many researchers may resist. In particular, there is a common sentiment that reviewer anonymity is advantageous, that it: protects social-professional relationships from anger aroused by criticism, allows for greater honesty since there is no concern about repercussions, and increases participation. However, in the current system the combination of anonymity, lack of accountability, and access to author material creates the potential for serious problems such as the use of the authors' ideas without acknowledgment of their source. Under the proposed system, people who implement the system will have the option to consider which components remain anonymous but reviewers would be tracked, potentially alleviating this issue. Furthermore, the open post-publication review system prevents any single person from blocking a publication or giving it a negative rating. The transparency of such a system will also reduce any single individual or group's ability to game the system. To further curtail the selfish tendencies of some reviewers, comments they make about the text would themselves be subject to review by others, and it would be in their own self-interest to maintain a high rating in their peer community.

In the long run, the review process should not be limited to publication, but should be engaged throughout the process of research, from inception through planning, execution, and documentation (Butler, 2005; see **Figure 1B**). Open review at every stage of a scientific study would facilitate collaborative research

and mirror open source project development closely. Such a process would also ensure that optimal decisions are taken at every stage in the evolution of a project, thus improving the quality of any scientific investigation. We envision a system where the distinction between authors and reviewers is replaced simply by a quantitative measure of contribution and scientific impact, especially as reviewers can act as collaborators who play a critical role in improving the quality and, therefore, the impact of scientific work. Where there is significant concern about exposing ideas before an article is written, reviewers could be drawn from collaborators, funding agencies, focus groups, or within the authors' institutions or laboratories, rather than the general public. In such scenarios either the review process or the identity of reviewers or both could be kept hidden but tracked for the purposes of "reputation assessment" (see above) and accountability.

Changing the review process in ways outlined in this article should lead to better science by turning each article into a public forum for scientific dialogue and debate. The proposed discussion-based environment will track and quantify impact of not only the original article, but of the comments made during the ensuing dialogue, helping readers to better filter, find, and follow this information while quantitatively acknowledging author and reviewer contributions and their quality. Our proposed re-design of the current peer review system focuses on post-publication review, and incorporates ideas from code review systems associated with open source software development. Such a system should enable a less biased, comprehensive, and efficient review of scientific work while ensuring a continued, evolving, public dialogue.

## ACKNOWLEDGMENTS

We would like to thank Matthew Goodman, Yaroslav Halchenko, Barrett Klein, Kim Lumbard, Fernando Perez, Jean-Baptiste Poline, Elizabeth Sublette, and the Frontiers reviewers for their helpful comments. Arno Klein would like to thank Deepanjana and Ellora, as well as the NIMH for their support via R01 grant MH084029. Brian Avants acknowledges ARRA funding from the National Library of Medicine via award HHSN276201000492p.

## REFERENCES

- Bollen, J., van de Sompel, H., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS One* 4:e6022. doi: 10.1371/journal.pone.0006022
- Butler, D. (2005). Electronic notebooks: a new leaf. *Nature* 436, 20–21.
- Chang, G., Roth, C. B., Reyes, C. L., Pornillos, O., Chen, Y.-J., and Chen, A. P. (2006). Retraction. *Science* 314, 1875.
- Drummond, C. (2009). "Replicability is not reproducibility: nor is it good science," in *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICMML*. (Montreal, Canada). Citeseer.
- Garfield, E. (1955). Citation indexes to science: a new dimension in documentation through association of ideas. *Science* 122, 108–111.
- Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* 102, 16569.
- Ioannidis, J. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124
- Leubsdorf, C. Jr. (2011). "Annotum: an open-source authoring and publishing platform based on WordPress," in *Proceedings of the Journal Article Tag Suite Conference*. (Bethesda, MD: National Center for Biotechnology Information US).
- Peng, R. D. (2009). Reproducible research and Biostatistics. *Biostatistics* 10, 405–408.
- Poschl, U., and Koop, T. (2008). Interactive open access publishing and collaborative peer review for improved scientific communication and quality assurance. *Inform. Serv. Use* 28, 105–107.
- Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *J. R. Soc. Med.* 99, 178.
- Viegas, F., Wattenberg, M., and Dave, K. (2004). "Studying cooperation and conflict between authors with history flow visualizations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (New York, NY, USA: ACM Press), 575–582.
- commercial or financial relationships that could be construed as a potential conflict of interest.

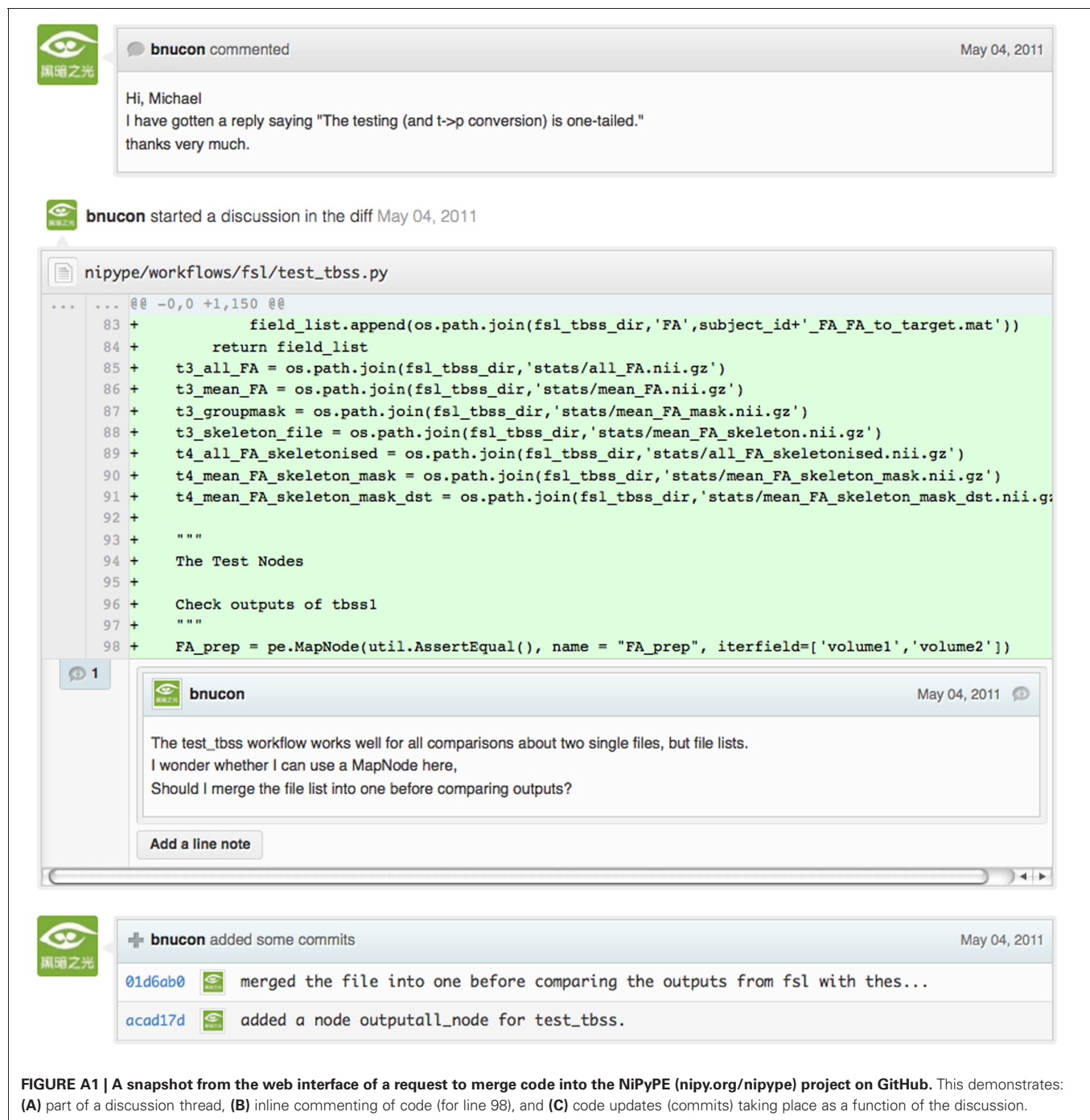
Received: 06 June 2011; accepted: 16 March 2012; published online: 18 April 2012.

Citation: Ghosh SS, Klein A, Avants B and Millman KJ (2012) Learning from open source software projects to improve scientific review. *Front. Comput. Neurosci.* 6:18. doi: 10.3389/fncom.2012.00018

Copyright © 2012 Ghosh, Klein, Avants and Millman. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



## APPENDIX



**FIGURE A1 |** A snapshot from the web interface of a request to merge code into the NiPyPE ([nipype.org/nipype](http://nipype.org/nipype)) project on GitHub. This demonstrates: (A) part of a discussion thread, (B) inline commenting of code (for line 98), and (C) code updates (commits) taking place as a function of the discussion.

Reviewer	Verified	Code Review	
<a href="#">Gaëtan Lehmann</a>			
<a href="#">Hans J. Johnson</a>		+1	Looks good to me, but someone else must approve
<a href="#">Andrew Wasem</a>		+1	Looks good to me, but someone else must approve
<a href="#">Jim Miller</a>			

- Need Verified +1 (Verified)
- Need Code Review +2 (Looks good to me, approved)

**FIGURE A2 | A web page snippet from the Geritt code review system used for Insight Toolkit (itk.org).** This explicitly lists the reviewers who are participating in the review.

## How does “Reputation” work?

▲ On Stack Exchange, users may gain a certain level of *reputation*.

170

▼

☆  
55

- What does *reputation* do?
- How can a user gain or lose *reputation*?

### 1 Answer

active oldest votes

▲

220

▼

#### What does *Reputation* do?

As a registered user, your reputation on the site is a part of your identity on the site. It determines, to an extent, your familiarity with the site, the amount of subject matter expertise you have and the level of respect your peers have for you. It can generally only be gained when other users of the site approve of the content you provide.

Reputation also determines a user's privileges within the system. As you gain more reputation, the system learns to trust you and bestows new functionality upon you that low-reputation users cannot access.

As users gain reputation, they gain abilities and responsibilities. The required reputation amounts on different sites can vary slightly; see [your site's /privileges page](#) for specifics. **Common privilege levels for new sites, public beta sites and "normal" sites are described [here](#).**

**FIGURE A3 | A response to a question on stackoverflow.net.** The top left number (170) indicates the number of positive votes this response received. There are comments to the response itself and the numbers next to the

comments reflect the number of positive votes for each comment (e.g., 220 in this example). (<http://meta.stackoverflow.com/questions/76251/how-do-suggested-edits-work>).





# Aggregating post-publication peer reviews and ratings

Răzvan V. Florian<sup>1,2,3\*</sup>

<sup>1</sup> Romanian Institute of Science and Technology, Cluj-Napoca, Romania

<sup>2</sup> Epistemio Ltd., London, UK

<sup>3</sup> Epistemio Systems SRL, Cluj-Napoca, Romania

## Edited by:

Diana Deca, Technical University  
Munich, Germany

## Reviewed by:

Dietrich S. Schwarzkopf, Wellcome  
Trust Centre for Neuroimaging at  
UCL, UK

Dwight Kravitz, National Institutes of  
Health, USA

## \*Correspondence:

Răzvan V. Florian, Romanian  
Institute of Science and Technology,  
Str. Ciresilor nr. 29, 400487  
Cluj-Napoca, Romania.  
e-mail: florian@rist.ro

Allocating funding for research often entails the review of the publications authored by a scientist or a group of scientists. For practical reasons, in many cases this review cannot be performed by a sufficient number of specialists in the core domain of the reviewed publications. In the meanwhile, each scientist reads thoroughly, on average, about 88 scientific articles per year, and the evaluative information that scientists can provide about these articles is currently lost. I suggest that aggregating in an online database reviews or ratings on the publications that scientists read anyhow can provide important information that can revolutionize the evaluation processes that support funding decisions. I also suggest that such aggregation of reviews can be encouraged by a system that would provide a publicly available review portfolio for each scientist, without prejudicing the anonymity of reviews. I provide some quantitative estimates on the number and distribution of reviews and ratings that can be obtained.

**Keywords:** peer review, post-publication peer review, scientific evaluation

## INTRODUCTION

There is an increasing awareness of the problems of the current scientific publication system, which is based on an outdated paradigm, resulted from the constraints of physical space in printed journals, and which largely ignores the possibilities opened by current internet technologies. There also is an increasing interest in alternatives to this paradigm (Greenbaum et al., 2003; Van de Sompel et al., 2004; Rodriguez et al., 2006; Carmi and Coch, 2007; Easton, 2007; Kriegeskorte, 2009; Chang and Aernoudts, 2010). Several papers within this journal's Research Topic on *Beyond open access* present convincingly a vision of a future where the scientific journal's functions are decoupled and/or the pre-publication reviews by about two or three reviewers is replaced or complemented by an ongoing post-publication process of transparent peer review and rating of papers (Kravitz and Baker, 2011; Lee, 2011; Ghosh et al., 2012; Priem and Hemminger, 2012; Sandewall, 2012; Wicherts et al., 2012; Zimmermann et al., 2012). This could ensure a better assessment of the validity of the information provided in a scientific publication, which would help those that intend to use that information in their research or in applications.

Peer review supports not just the scientific publication system, but also the allocation of funding to scientists and their institutions. Just as an open post-publication peer review process can revolutionize scientific publication, it can also revolutionize the evaluation procedures that support funding decisions. I will argue here that aggregating post-publication peer reviews is a better alternative to organizing dedicated review committees and I will suggest some mechanisms to motivate the aggregation of such peer reviews.

## PEER REVIEW FOR FUNDING DECISIONS

Funding decisions include: funding research projects through grants; allocating funding to a research group or an institution; and hiring or granting tenure. These decisions are typically based, in a significant measure, on a review by a committee of the previous results of a scientist or of a group of scientists, and in many cases these results are scientific publications.

In many cases, because of practical issues, the review committee does not include specialists in the core area of expertise of the assessed scientists. Such practical issues include:

- selecting reviewers from in-house databases that are not comprehensive or up-to-date, which limits the range of potential reviewers to those in the database;
- selecting reviewers using software that matches them to assessed scientists using keywords or matching between broad domains, a method that can lead to imprecise results; better results could be obtained if matching would be based on co-authorship networks (Rodriguez and Bollen, 2008) or co-citation networks;
- the lack of time or other reasons for the unavailability of selected reviewers, especially when the type of review requires a trip or other significant time investment from the reviewers.

Because of the increased specialization of modern science, this can prevent a thorough understanding by the reviewer of the assessed scientist's publications. In other cases, the reviewers simply do not have the time to thoroughly read and properly assess these publications. These situations may lead the reviewers to rely on indirect, more imprecise indicators of the publication's quality,

such as the impact factor of the journal where it was published, its number of citations or other such metrics, instead of the publication's content, as they should do, and thus may lead to a higher subjectivity of the review.

For example, the Research Assessment Exercise (RAE) has been used in the UK for allocating funding to universities, of about £1.5 billion per year (HEFCE, 2009). A key part of the exercise was the peer review of outputs, typically publications, submitted by universities. Universities were allowed to submit up to four output items per each of the selected university staff members. About 1400 members of the RAE review panels reviewed 214,287 outputs<sup>1</sup>, i.e., on average there were about 150 outputs per reviewer. This large number of outputs that a reviewer had to assess means that only few of them were thoroughly reviewed.

Even minor improvements in evaluation of science, that would improve the efficiency of the allocation of research funding, would translate in huge efficiency increases, as the global research and development spending is about 1143 billion US dollars annually (Advantage Business Media, 2009). For example, a 1% relative improvement would lead to worldwide efficiency increases of about 11 billion dollars annually.

## AGGREGATING POST-PUBLICATION PEER REVIEWS AND RATINGS

An alternative to reviewing publications independently for each funding allocation decision is centralizing and aggregating reviews or ratings from each scientist who reads the papers for her own needs. Internet technologies make quite simple to implement a system where reviews or ratings collected through one or multiple websites or mobile applications are centralized in a single database. Encouraging a simple procedure, that a scientist goes to a website or opens a mobile app and spends several minutes logging there her rating or review information on each new publication that she reads, would provide a much more precise and relevant review information than most of the currently available processes. In many cases, this would entail just collecting existing information, the evaluative valences of which are otherwise wasted for the society.

Scientists spend anyhow a large percentage of their time reading scientific publications—the results of two studies point to 6% or, respectively, 38% of the work time as being spent reading, on average (Tenopir et al., 2009, 2011), assuming a 45 h work week (Table 1). A questionnaire performed on US university staff in 2005 has shown that a scientist reads, on average, 204 unique articles per year, for a total of 280 readings. The average reading time was 31 min (Tenopir et al., 2009). Forty-three percent of these readings (i.e., about 88 unique articles per year) were read “with great care” and 51% were read “with attention to the main points” (Tenopir et al., 2009) (see also Table 2). While reading a paper, scientists form an opinion about its quality and relevance, and this opinion could be collected by an online service as a rating of the paper. Across the world, journal clubs are organized periodically in most universities and research institutes, where scientists discuss new publications. Again, the results of these

discussions could be collected by an online service, as reviews of those publications.

There is a quite large gap between the average number of articles that a scientist reads with great care (88 per year) (Tenopir et al., 2009) and the average number of articles that a scientist reviews (8 per year) (Ware and Monkman, 2008). Review information on the about 80 articles per scientist per year that were not specifically read for review is currently lost.

The people that would provide these ratings and reviews are typically specialists in the core field of the publications they review, unlike many of the reviewers in committees formed for decision making. If this information would be aggregated globally, from all scientists in the world who read a particular publication, the accuracy and relevance of the review information would be much higher than the one available through classical means.

This review information might be similar in scope to the one provided in typical pre-publication reviews. However, brief reviews or just ratings of the scientific articles on a few dimensions would also be informative when many of them (e.g., 10 or more) would be aggregated. As I discuss below, we can expect that only a fraction of publications will get, e.g., three or more reviews or 10 or more ratings.

The content of the reviews would be made public. Reviews could be rated themselves, and this would provide information from scientists that do not have the time to write the reviews themselves but just to express their agreement or disagreement with existing reviews. The rating of reviews would also encourage their authors to pay attention to the quality of these reviews (Wicherts et al., 2012).

Although the reviews or ratings could be kept anonymous for the public if their authors desire it, the identity of the reviewers should be checked by the providers of the proposed system in order to ensure the relevance of the aggregated information. The relevance of the reviews and ratings could be weighted by the scientific prestige of their authors and by the fit between the reviewer's and the reviewed paper's fields. This scientific prestige could be assessed initially using classical scientometric indicators, but once reviews and ratings would start being aggregated these would be used increasingly for assessing scientific prestige. Synthetic indicators of scientific prestige built upon the aggregated review information should take into account differences between different fields of research in publication frequency and impact. These synthetic indicators should also be presented with error bars/confidence intervals (Kriegeskorte, 2009) or as distributions and not only just as unique numerical values, like current scientometric indicators are typically presented.

One problem that is often mentioned about the present system of pre-publication review is the issue of political reviewing—unjustified negative reviews of papers of direct competitors or of scientists supporting competing views (Smith, 2006; Benos et al., 2007). Because the pre-publication review typically leads to a binary decision (accept or reject the publication), one negative, unjust review by a competitor can lead to a negative outcome even if other reviews are positive. Since the proposed system would also consider and display the distribution of reviews/ratings, a paper that is highly acclaimed by a significant percentage of reviewers could be considered as an interesting one even if another

<sup>1</sup> <http://www.rae.ac.uk/>

**Table 1 | Total work time and time spent on various tasks.**

Task	Average time spent (hours per week)	Population	Reference
Total work time	48.52	Worldwide highly cited scientists in environmental science and ecology	Parker et al., 2010
Total work time	52	Doctoral level academics in biological and agricultural sciences	Parker et al., 2010
Total work time	39.3	European active population, 2009	Carley, 2010
Reading scientific articles	2.78	US university staff, 2005	Tenopir et al., 2009
Reading scientific articles	17.25	Academic staff members at 7 universities in 7 countries, 2008	Tenopir et al., 2011
Reviewing publications	1.30	Typical scientists	Ware and Monkman, 2008
Reviewing publications	1.86	Active reviewers	Ware and Monkman, 2008
Reviewing manuscripts and grants	5.02	Worldwide highly cited scientists in environmental science and ecology	Parker et al., 2010
Spending 1 h each month for writing a review for an already read publication	0.25	Scientists	Direct computation
Spending 10 min each week for adding on a website a rating for an already read publication	0.17	Scientists	Direct computation

**Table 2 | The average number per year of readings, reviews and related activities that a scientist performs.**

Items	Average number per year	Population	Reference
Articles read or re-read	150	US university staff, 1977	Tenopir et al., 2009
Articles read or re-read	280	US university staff, 2005	Tenopir et al., 2009
Articles read or re-read	414	US medical faculty, 2005	Tenopir et al., 2009
Articles read or re-read	331	US science faculty, 2005	Tenopir et al., 2009
Articles read or re-read	223	US social sciences faculty, 2005	Tenopir et al., 2009
Unique articles read	204	US university staff, 2005	Tenopir et al., 2009
Unique articles read with great care	88	US university staff, 2005	Tenopir et al., 2009
Unique articles read with attention to the main points	104	US university staff, 2005	Tenopir et al., 2009
Articles reviewed	8.0	Typical scientists	Ware and Monkman, 2008
Articles reviewed	14.3	Active reviewers	Ware and Monkman, 2008
Articles that scientists are prepared to review	9.0	Typical scientists	Ware and Monkman, 2008

significant percentage of reviewers consider it in a negative way. Automated mechanisms could easily be developed to distinguish between unimodal and bimodal distributions of ratings. Simple checks based on coauthorship information or institutional affiliations and more complex checks based on the detection of citation circles or reciprocal reviewing could also filter out other more general conflicts of interest (Aleman-Meza et al., 2006).

Enabling the collection of reviews and ratings through mobile applications is important, since only 64.7% of article readings happen in the office or lab, while 25.7% happen at home and 4.1% while traveling, on average (Tenopir et al., 2009).

## THE CURRENT STATUS OF REVIEW AGGREGATION

There were many attempts to collect post-publication review information, but, to date, despite the enthusiasm for the concept, the number of reviews provided through the available channels

is deceptively low. For example, the prestigious journal *Nature* launched a trial of open peer review, which proved to be not widely popular, either among authors or by scientists invited to comment (Greaves et al., 2006). Some of the causes of this outcome could have been corrected, however (Pöschl, 2010). PLoS ONE<sup>2</sup> peer reviews submissions on the basis of scientific rigor, leaving the assessment of the value or significance of any particular article to the post-publication phase (Patterson, 2010). So far, the usage of the commentary tools of PLoS ONE is fairly modest and does not make a major contribution to the assessment of research content (Public Library of Science, 2011). Innovative journals such as Philica<sup>3</sup> or WebMedCentral<sup>4</sup>

<sup>2</sup><http://www.plosone.org/>

<sup>3</sup><http://philica.com/>

<sup>4</sup><http://webmedcentral.com/>

that aim to provide only post-publication review for the papers that they publish suffer from a lack of reviews and are overwhelmed by low quality papers. Faculty of 1000<sup>5</sup> organizes the review of about 1500 articles monthly, corresponding to approximately the top 2% of all published articles in the biology and medical sciences<sup>6</sup>, but this covers just a few scientific areas, a small fraction of the publications within these areas, and accepts reviews from a limited pool of scientists only. The Electronic Transactions on Artificial Intelligence<sup>7</sup>, which combined open post-publication review with a traditional accept/reject decision by editor-appointed reviewers, seems to be an example of moderate success (Sandewall, 2012), but is, however, currently closed. The Atmospheric Chemistry and Physics<sup>8</sup> journal and its sister journals of the European Geosciences Union and Copernicus Publications, which use an interactive open access peer review, also are examples of moderate success, but only about 25% of the papers receive a comment from the scientific community in addition to the comments from designated reviewers, for a total average of about 4–5 interactive comments (Pöschl, 2010).

All these show that the existing mechanisms and incentives are not sufficient to encourage scientists to contribute a significant number of reviews.

Scientists are quite busy and work long hours, the work time being about 30% higher than the average one of the general population (Table 1). Over the last few decades, as the number of scientific publications and their accessibility has grown, the average number of articles read by scientists has increased from 150 per year in 1977 to 280 per year in 2005. However, the average time spent reading a paper has decreased from 48 to 31 min, suggesting that the amount of time available for reading scientific articles is likely reaching a maximum capacity (Tenopir et al., 2009).

The highly cited scientists in environmental science and ecology spend, on average, more than 10% of their work time reviewing manuscripts and grants (Parker et al., 2010). Typical scientists review, on average, 8.0 papers per year, which takes, on average, 8.5 h per paper (median 5 h) (Ware and Monkman, 2008). Active reviewers review, on average, 14.3 papers per year, for 6.8 h per paper (Ware and Monkman, 2008) (see also Tables 1 and 2). The average number of papers per year that scientists are prepared to review is 9.0, i.e., slightly higher than the 8.0 papers they actually review, however, the active reviewers are overloaded (Ware and Monkman, 2008). The lack of time is a major factor determining the decision to decline to review a paper (Table 3).

All these suggest that it is not realistic to expect that scientists can spend significantly more of their time reading new articles just for the purpose of providing reviews for them, unless there would be some strong incentives for doing so. However, logging on a website review or rating information for some of the articles that they have already read would not be a significant burden, as estimated below.

<sup>5</sup><http://f1000.com/>

<sup>6</sup><http://f1000.com/thefaculty>

<sup>7</sup><http://www.etaij.org/>

<sup>8</sup><http://www.atmospheric-chemistry-and-physics.net/>

**Table 3 | Most important factors in the decision to decline to review a paper.**

Factor	Reference
Conflict with other workload; a tight deadline for completing the review; having too many reviews for other journals	Tite and Schroter, 2007
Lack of expertise in the paper's domain; lack of time	Lu, 2008
The paper was outside the scientist's area of expertise; the scientist was too busy doing her own research, lecturing, etc.; too many prior reviewing commitments	Sense About Science, 2009

In a survey (Schroter et al., 2010), 48% of scientists said their institution or managers encouraged them to take part in science grant review, yet only 14% said their institution or managers knew how much time they spent reviewing and 31% knew what funding organizations they reviewed for. A total of 32% were expected to review grants in their own time (out of office hours) and only 7% were given protected time to conduct grant review. A total of 74% did not receive any academic recognition for conducting grant review (Schroter et al., 2010). This suggests that, currently, institutions do not reward sufficiently the scientists' review activities.

## PREVIOUS SUGGESTIONS FOR ENCOURAGING THE AGGREGATION OF REVIEWS

Several surveys asked reviewers about their motivation to review and the factors that would make them more likely to review. The main motivations for reviewing are: playing one's part as a member of the academic community; enjoying being able to help improve the paper; enjoying seeing new work ahead of publication; reciprocating the benefit gained when others review your papers (Ware and Monkman, 2008; Sense About Science, 2009). The incentives that would best encourage reviewers to accept requests to review are presented in Table 4.

A potential reviewer's decision to spend time reviewing an article, which yields an information that is a public good, as opposed to the alternative of spending time in a way that is more directly beneficial to the reviewer, can be construed as a social dilemma. Northcraft and Tenbrunsel (2011) argue that reviewers' cooperation in this social dilemma depends on the costs and the benefits as personally perceived by the reviewers. This personal perception may be influenced by the frame reviewers bring to the decision to review. Frames may lead reviewing to be viewed as an in-role duty or an extra-role choice, and may lead reviewers to focus only on consequences to the self or consequences to others as well (Northcraft and Tenbrunsel, 2011). This theoretical framework allowed Northcraft and Tenbrunsel to suggest several methods for improving cooperation within this social dilemma, among which are:

- institutions that employ the reviewers should encourage the perspective that reviewing is an in-role duty, by recognizing



**Table 4 | The most important factors that would encourage scientists to review papers.**

Factor	Reference
Free access or subscription to journal content; annual acknowledgement on the journal's website; more feedback about the outcome of the submission and quality of the review; appointment of reviewers to the journal's editorial board; published acknowledgement of reviewer's contribution to the manuscript; consultancy-equivalent fee for time spent; small financial contributions, e.g., lower than £50	Tite and Schroter, 2007
Free subscription to the journal; acknowledgement in the journal (e.g., appear in the list of most frequent reviewers); payment in kind by the journal (e.g., waiver of color or other publication charges, free offprints, etc.); optional accreditation for CME/CPD points (mainly of interest to clinical researchers)	Ware and Monkman, 2008
Payment in kind by the journal; payment by the journal; acknowledgement in the journal; accreditation (CME/CPD points). While 41% of respondents would be incentivized by receiving payment for reviewing, the percentage drops to 2.5% if the author had to cover the cost	Sense About Science, 2009

and rewarding reviewing in evaluations of reviewers' professional activity;

- creating a public database of reviewers, which would increase reviewer accountability by communicating publicly who is and who is not reviewing, and thus decreasing the probability of undetected free riding.

Another suggestion to discourage free riding in the reviewer social dilemma was to establish a credit system to be used by all journals, where a scientist's account would be credited for his/her reviews and debited when he/she submits a paper for review (Fox and Petchey, 2010).

Other suggestions include considering reviews as citable publications in their own right, which will motivate reviewers in terms of quality and quantity (Kriegeskorte, 2009).

Another option would be to simply eliminate the dilemma, by considering that reviewers should read and review only papers that are of direct interest to them to read, and by considering that the papers that are not read and reviewed are simply not worthy of attention and presumably of low quality (Lee, 2011).

## MOTIVATING THE AGGREGATION OF REVIEWS AND RATINGS

I propose a simple system that aims to motivate the aggregation of reviews and ratings by reinforcing the in-role duty of the reviewers, by recognizing publicly that by reviewing they play their part as a member of the academic community, and by facilitating the reward by their institutions of their review and rating activities. Critically, this system would do this without prejudicing the anonymity of reviews, an issue that reviewers are quite keen about (Sense About Science, 2009).

The proposed system will build a review and rating portfolio for each scientist, which would be publicly available, similar to the publication or citation portfolios of scientists, which are currently used to reward them. The system would need a mechanism for uniquely identifying scientists, which hopefully will be provided soon by the Open Researcher and Contributor ID (ORCID)<sup>9</sup>. Each journal or grant giving agency, once authenticated, will be able to register to the system the identity of the reviewers that helped them and, possibly, to rate the reviewer's contribution.

This information provided by the journals or the agencies, i.e., a quantity representing the extent of the reviewer's contribution and another quantity representing the quality of the reviewer's contribution, will be made public after a random timing. This random timing will be chosen such that it will not be possible for the public, including the reviewed scientists, to associate the change of the reviewer's public information to the actual review, and thus to establish the identity of the reviewers who performed a given review. The anonymity of reviews will thus be respected.

Once there will be a system that will provide this kind of information, in a certified manner (with the contribution of journals and funding agencies), it will be easier for institutions to reward reviewing. As presented above, institutions do not reward sufficiently the in-role duty of scientists to review. A possible cause for this is the lack of easy access to information about a scientist's contribution to peer review, certified by a third party other than the scientist. The proposed system will provide this information, thus facilitating institutions to reward reviews and, finally, contributing to a higher participation of scientists to peer review.

This system can be then extended to account not only for pre-publication reviews and the review of grant applications, but also for post-publication reviews and ratings. For ratings, the portfolio would include their number. Public reviews, such as post-publication reviews, could be rated themselves by others, and thus public information on the review quality of a particular scientist could be made available (Wicherts et al., 2012). Highly rated reviews could then be published as independent publications. Such a system is currently being developed by Epistemio<sup>10</sup>.

However, it is likely that scientists will not spend time reading papers that would not interest them. A large proportion of scientific publications is not cited and probably not read by scientists other than the authors and the reviewers involved in the publication, and thus, there will always be a significant percentage of papers that would not attract post-publication peer reviews. Scientists prefer reading papers written by an author they recognize as a top scholar and published in a top-tier peer-reviewed journal (Tenopir et al., 2010). Thus, it would be a challenge for young or emerging authors publishing in middle- or low-tier journals to attract the attention of relevant reviewers, even in

<sup>9</sup><http://orcid.org/>

<sup>10</sup><http://epistemio.com/>



the case that their results are important. If post-publication peer review will gain importance in supporting funding decisions or as a complement or replacement of pre-publication peer review, then the interested parties will have the option of offering incentives, including direct payment, to competent reviewers to spend their time reading and reviewing articles that did not attract initially the attention of other scientists. If the quality of these papers will be mostly low and the process will ensure the independence and the competence of reviewers, then the result of the process will reflect this quality. However, there are chances that this process would sort out a small proportion of important papers within the ones that did not attract attention initially, and this would motivate the process.

### ESTIMATING THE DISTRIBUTION OF THE NUMBER OF POST-PUBLICATION REVIEWS PER ARTICLE

Throughout this section, we consider that review means either a proper review or a rating, where not distinguished explicitly. Let's consider that the population of scientists who write papers is identical to the population of scientists who read scientific papers, and this population consists of  $N$  scientists. Let's consider that each of the scientists writes, on average, two full articles per year, where a scientist's contribution to a multi-author paper is accounted for fractionally (Tenopir and King, 1997). This means that all scientists write  $2N$  articles per year. If a scientist reads with great care about 88 unique articles per year, on average (Tenopir et al., 2009), this means that, if all scientists would log reviews for all these articles read with great care, there would be  $88N$  reviews per year. This leads to an average of about  $88N/2N = 44$  aggregated reviews per article. In practice, a fraction of scientists would log reviews for a fraction of the articles they read, and the average number or reviews per article will be lower than 44.

Scientometric distributions are much skewed: few articles attract a lot of attention and most of the articles attract little attention, and thus, the actual number of reviews per article will be in many cases far from the average. Let's assume that the number of reviews that an article attracts is proportional to the number of citations it attracts. A previous study has found that the dependence on the number of citations  $c$  of the number of articles that are cited  $c$  times can be fitted well, except for large numbers of citations, by a stretched exponential (Redner, 1998). For the general scientific literature, this exponential coefficient was  $\beta = 0.44$ . I will use here a simple model where I consider a continuous probability density  $p(x)$  for the number  $x$  of reviews that an article has. I consider that this probability density is such a stretched exponential,

$$p(x) = \frac{\Gamma(2/\beta)}{a \Gamma(1/\beta) \Gamma(1 + 1/\beta)} \exp \left[ - \left( \frac{x \Gamma(2/\beta)}{a \Gamma(1/\beta)} \right)^\beta \right],$$

where  $\Gamma$  is the Gamma function and  $a$  is a positive parameter. The form of  $p$  is chosen such that the average number of reviews per article is  $a$ ,

$$\int_0^\infty x p(x) dx = a.$$

For an integer number of reviews  $n$ , the fraction of the articles having  $n$  reviews can be approximated as

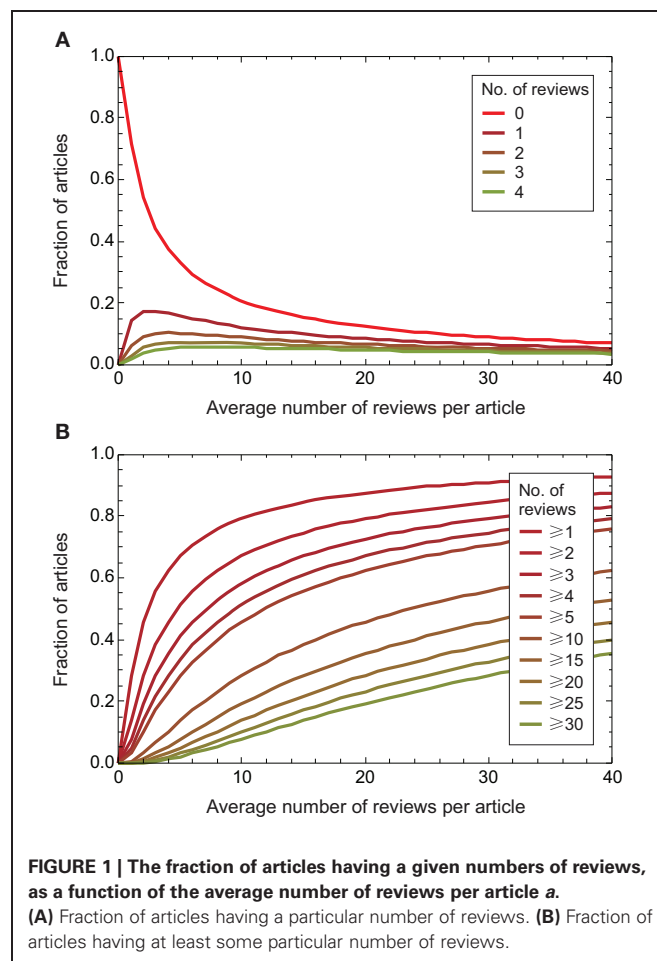
$$q(n) = \int_n^{n+1} p(x) dx.$$

Figure 1 shows the fraction of the articles that get given numbers of reviews, as a function of the average number or reviews per article  $a$ , under this model.

Under a moderately optimistic example, let's assume that each scientist logs, on average, one proper review monthly and one rating weekly. Thus, there would be  $12N/2N = 6$  proper reviews per article, on average, and  $52N/2N = 26$  ratings per article, on average. The distribution of the number of proper reviews and ratings per article is presented in Table 5. Forty-six percent of the articles would get three or more proper reviews, and 52% of the articles would get 10 or more ratings, thus receiving enough evaluative information for a proper assessment of the article's relevance.

### ESTIMATING THE ADDITIONAL TIME BURDEN ON SCIENTISTS OF POST-PUBLICATION REVIEWS

As mentioned above, currently a pre-publication review for a paper takes, on average, 8.5 h (median 5 h) for a typical scientist,



**FIGURE 1 | The fraction of articles having a given numbers of reviews, as a function of the average number of reviews per article  $a$ .** (A) Fraction of articles having a particular number of reviews. (B) Fraction of articles having at least some particular number of reviews.

**Table 5 | An estimated distribution of the number of proper reviews and ratings per article, assuming that each scientist logs, on average, one proper review monthly and one rating weekly.**

Proper reviews ( $a = 6$ )	
Number of reviews	Percentage (%)
0	29
1...2	25
3...4	13
5...9	16
≥10	17
Ratings ( $a = 26$ )	
Number of ratings	Percentage (%)
0	10
1...4	22
5...9	16
10...19	17
20...29	10
≥30	25

and 6.8 h for an active reviewer (Ware and Monkman, 2008). This probably includes the time needed for reading the paper, reading additional relevant papers cited in the reviewed paper, and actually conceiving and writing the review.

We focused here on the aggregation of reviews and ratings of publications that are read anyhow by the scientists, so the time needed for reading the reviewed publications is not an additional burden in our case. It is also possible that in many cases reviewers for papers currently submitted for publication receive and accept for review papers that are not in their core field of research, hence the possible need for an extra documentation requiring reading some of the publications cited in the reviewed paper. As already mentioned above, the lack of expertise in the paper's domain is an often mentioned reason for refusing a review (Lu, 2008; Sense About Science, 2009), which means that receiving for review papers that are not in the reviewer's core field of research is common. In the case of reviews of papers that scientists read anyhow, these papers are guaranteed to be from their core field, and thus, reading extra publications cited in the reviewed papers is not an additional burden.

Thus, the time needed for reading the actual paper and any additional papers must be subtracted from the current duration

of pre-publication reviews in order to estimate the time spent for just conceiving and writing a review. The average reading time for an article is 31 min (Tenopir et al., 2009), but this averages the time spent for reading articles with various degrees of attention. Thus, we would expect that the time reading an article with great care is somehow larger than 31 min, on average. After subtracting the time needed for reading the papers, a reasonable estimate of the time spent for just conceiving and writing a proper review for an already read paper is of about 1 h.

The time needed to access a website or a mobile app, search for the publication that has just been read and add ratings on a few dimensions can also be reasonably estimated to about 10 min.

The additional time burden resulted from these estimates, for logging, on average, one proper review monthly, and one rating weekly, is presented in **Table 2** together with the time burden of other activities and appears to be small.

This extra work will be later compensated by less time spent on searching for relevant information, when review information will be available to filter articles of interest. In the cases where post-publication peer review will replace the pre-publication peer review, there would be no extra work at all.

## CONCLUSIONS

I have argued that the aggregation of post-publication peer reviews and ratings can play an important role for revolutionizing not only scientific publication, but also the evaluation procedures that support funding decisions. I have presented some suggestions for motivating scientists to log such reviews or ratings. I have also estimated quantitatively the maximum average number of reviews/ratings and the distribution of the number of reviews/ratings that articles are expected to receive if reviews/ratings for some of the articles that scientists read thoroughly are logged online and centralized in a database.

The internet has revolutionized many aspects of economy and society, such as communication, press, travel, music, and retail. Although the scientific enterprise is centered around information, it resisted to date to a significant embrace of the possibilities of online collaboration and information sharing offered by the internet. Besides moving the publications from print to web and allowing an easier access to publications, the advent of the internet has not changed much the scientific enterprise. A centralized aggregation of reviews and ratings of scientific publications can provide better means to evaluate scientists, thus allowing improved efficiencies in allocating research funding and accelerating the scientific process.

## REFERENCES

- Advantage Business Media (2009). 2009 Global R&D funding forecast.
- Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A., Arpinar, I., Joshi, A., and Finin, T. (2006). "Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection," in *Proceedings of the 15th International Conference on World Wide Web* (New York, NY: ACM), 407–416.
- Benos, D. J., Bashari, E., Chaves, J. M., Gaggari, A., Kapoor, N., LaFrance, M., Mans, R., Mayhew, D., McGowan, S., Polter, A., Qadri, Y., Sarfare, S., Schultz, K., Splittgerber, R., Stephenson, J., Tower, C., Walton, R. G., and Zotov, A. (2007). The ups and downs of peer review. *Adv. Physiol. Educ.* 31, 145–152.
- Carley, M. (2010). Working time developments – 2009. Available from <http://www.eurofound.europa.eu/eiro/studies/tn1004039s/tn1004039s.htm>
- Carmi, R., and Coch, K. (2007). Improving peer review with CARMA. *Learn. Publ.* 20, 173–176.
- Chang, C.-M., and Aernoudts, R. H. R. M. (2010). Towards scholarly communication 2.0, peer-to-peer review and ranking in open access preprint repositories. Available from SSRN: <http://ssrn.com/abstract=1681478>
- Easton, G. (2007). Liberating the markets for journal publications: some specific options. *J. Manage. Stud.* 44, 4.
- Fox, J., and Petchey, O. (2010). Pubcres: fixing the peer review process by "privatizing" the reviewer commons. *Bull. Ecol. Soc. Am.* 91, 325–333.
- Ghosh, S. S., Klein, A., Avants, B., and Millman, K. J. (2012). Learning from open source software projects to improve scientific review. *Front. Comput. Neurosci.* 6:18. doi: 10.3389/fncom.2012.00018

- Greaves, S., Scott, J., Clarke, M., Miller, L., Hannay, T., Thomas, A., and Campbell, P. (2006). Nature's trial of open peer review. *Nature*. doi: 10.1038/nature05535
- Greenbaum, D., Lim, J., and Gerstein, M. (2003). An analysis of the present system of scientific publishing: what's wrong and where to go from here. *Interdiscipl. Sci. Rev.* 28, 293–302.
- HEFCE (2009). Research excellence framework: second consultation on the assessment and funding of research. Available from [http://www.hefce.ac.uk/pubs/hefce/2009/09\\_38/09\\_38.pdf](http://www.hefce.ac.uk/pubs/hefce/2009/09_38/09_38.pdf)
- Kravitz, D., and Baker, C. I. (2011). Toward a new model of scientific publishing: discussion and a proposal. *Front. Comput. Neurosci.* 5:55. doi: 10.3389/fncom.2011.00055
- Kriegeskorte, N. (2009). The future of scientific publishing: ideas for an open, transparent, independent system. Available from <http://futureofscipub.wordpress.com>
- Lee, C. (2011). Open peer review by a selected-papers network. *Front. Comput. Neurosci.* 6:1. doi: 10.3389/fncom.2012.00001
- Lu, Y. (2008). Peer review and its contribution to manuscript quality: an Australian perspective. *Learn. Publ.* 21, 307–318.
- Northcraft, G. B., and Tenbrunsel, A. E. (2011). Effective matrices, decision frames, and cooperation in volunteer dilemmas: a theoretical perspective on academic peer review. *Organ. Sci.* 22, 1277–1285.
- Parker, J. N., Lortie, C., and Allesina, S. (2010). Characterizing a scientific elite: the social characteristics of the most highly cited scientists in environmental science and ecology. *Scientometrics* 85, 129–143.
- Patterson, M. (2010). PLoS ONE: Editors, contents and goals. Available from <http://blogs.plos.org/plos/2010/05/plos-one-editors-contents-and-goals>
- Priem, J., and Hemminger, B. H. (2012). Decoupling the scholarly journal. *Front. Comput. Neurosci.* 6:19. doi: 10.3389/fncom.2012.00019
- Public Library of Science (2011). "Peer review—optimizing practices for online scholarly communication," in *Peer Review in Scientific Publications, Eighth Report of Session 2010–2012, Vol. I: Report, Together with Formal, Minutes, Oral and Written Evidence*, eds House of Commons Science and Technology Committee (London: The Stationery Office Limited), Ev 77–Ev 81.
- Pöschl, U. (2010). Interactive open access publishing and peer review: the effectiveness and perspectives of transparency and self-regulation in scientific communication and evaluation. *Liber Q.* 19, 293–314.
- Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B* 4, 131–138.
- Rodriguez, M., and Bollen, J. (2008). "An algorithm to determine peer-reviewers," in *Proceeding of the 17th ACM Conference on Information and Knowledge Management* (New York, NY: ACM), 319–328.
- Rodriguez, M. A., Bollen, J., and Van de Sompel, H. (2006). The convergence of digital libraries and the peer-review process. *J. Inf. Sci.* 32, 149–159.
- Sandewall, E. (2012). Maintaining live discussion in two-stage open peer review. *Front. Comput. Neurosci.* 6:9. doi: 10.3389/fncom.2012.00009
- Schroter, S., Groves, T., and Højgaard, L. (2010). Surveys of current status in biomedical science grant review: funding organisations' and grant reviewers' perspectives. *BMC Med.* 8, 62.
- Sense About Science (2009). Peer review survey 2009. Available from <http://www.senseaboutscience.org/pages/peer-review-survey-2009.html>
- Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *J. R. Soc. Med.* 99, 178–182.
- Tenopir, C., Allard, S., Bates, B., Levine, K. J., King, D. W., Birch, B., Mays, R., and Caldwell, C. (2010). Research publication characteristics and their relative values: a report for the publishing research consortium. Available from <http://www.publishingresearch.org.uk/documents/PRCReportTenopiretalJan2011.pdf>
- Tenopir, C., and King, D. W. (1997). Trends in scientific scholarly journal publishing in the United States. *J. Sch. Publ.* 28, 135–170.
- Tenopir, C., King, D., Edwards, S., and Wu, L. (2009). "Electronic journals and changes in scholarly article seeking and reading patterns," in *Aslib Proceedings: New Information Perspectives*, Vol. 61, (Bingley, UK: Emerald Group Publishing Limited), 5–32.
- Tenopir, C., Mays, R., and Wu, L. (2011). Journal article growth and reading patterns. *New Rev. Inf. Netw.* 16, 4–22.
- Tite, L., and Schroter, S. (2007). Why do peer reviewers decline to review? A survey. *J. Epidemiol. Community Health* 61, 9–12.
- Van de Sompel, H., Erickson, J., Payette, S., Lagoze, C., and Warner, S. (2004). Rethinking scholarly communication: building the system that scholars deserve. *D-Lib Magazine* 10, 9.
- Ware, M., and Monkman, M. (2008). Peer review in scholarly journals: perspective of the scholarly community – an international study. Available from <http://www.publishingresearch.net/documents/PeerReviewFullPRCReport-final.pdf>
- Wichert, J. M., Kievit, R. A., Bakker, M., and Borsboom, D. (2012). Letting the daylight in: reviewing the reviewers and other ways to maximize transparency in science. *Front. Comput. Neurosci.* 6:20. doi: 10.3389/fncom.2012.00020
- Zimmermann, J., Roebroek, A., Uludag, K., Sack, A. T., Formisano, E., Jansma, B., Weerd, P. D., and Goebel, R. (2012). Network-based statistics for a community driven transparent publication process. *Front. Comput. Neurosci.* 6:11. doi: 10.3389/fncom.2012.00011

**Conflict of Interest Statement:** The author has financial interests in the Epistemo group of companies, which aim to provide commercial services related to aggregation of post-publication peer reviews and ratings.

Received: 16 February 2012; paper pending published: 07 March 2012; accepted: 07 May 2012; published online: 22 May 2012.

Citation: Florian RV (2012) Aggregating post-publication peer reviews and ratings. *Front. Comput. Neurosci.* 6:31. doi: 10.3389/fncom.2012.00031

Copyright © 2012 Florian. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.



# FOSE: a framework for open science evaluation

Alexander Walther<sup>1\*</sup> and Jasper J. F. van den Bosch<sup>2</sup>

<sup>1</sup> Cognition and Brain Sciences Unit, Medical Research Council, Cambridge, UK

<sup>2</sup> Institute of Medical Psychology, Goethe-University Frankfurt, Frankfurt am Main, Germany

## Edited by:

Nikolaus Kriegeskorte, Medical Research Council Cognition and Brain Sciences Unit, UK

## Reviewed by:

Satrajit S. Ghosh, Massachusetts Institute of Technology, USA  
Jason Priem, University of North Carolina at Chapel Hill, USA

## \*Correspondence:

Alexander Walther, Cognition and Brain Sciences Unit, Medical Research Council, 15 Chaucer Road, Cambridge CB2 7EF, UK.  
e-mail: alexander.walther@mrc-cbu.cam.ac.uk

Pre-publication peer review of scientific literature in its present state suffers from a lack of evaluation validity and transparency to the community. Inspired by social networks, we propose a framework for the open exchange of post-publication evaluation to complement the current system. We first formulate a number of necessary conditions that should be met by any design dedicated to perform open scientific evaluation. To introduce our framework, we provide a basic data standard and communication protocol. We argue for the superiority of a provider-independent framework, over a few isolated implementations, which allows the collection and analysis of open evaluation content across a wide range of diverse providers like scientific journals, research institutions, social networks, publishers websites, and more. Furthermore, we describe how its technical implementation can be achieved by using existing web standards and technology. Finally, we illustrate this with a set of examples and discuss further potential.

**Keywords:** open evaluation, peer review, social networking, standard

## INTRODUCTION

The success of scientific ideas critically depends on their successful publication. An unpublished idea, innovative, and promising as it might be, remains just that; only after publication it becomes a legitimate part of the scientific consciousness. A central gatekeeper function between the multiplicity of ideas and their manifestation as scientific publications is assigned to formal reviews governed by scientific journals. The current publishing system hinges on voluntary pre-publication peer review, with reviewers selected by the editorial staff. Peer review is undeniably a vital means of research evaluation for it is based on mutual exchange of expertise. Its role in the current system, however, has been the subject of concern with regard to accuracy, fairness, efficiency, and the ability to assess the long-term impact of a publication for the scientific community (Casati et al., 2010). For instance, studies suggest that peer review does not significantly improve manuscript quality (Goodman et al., 1994; Godlee et al., 1998) and that it is susceptible to biases to affiliation (Peters and Ceci, 1982) and gender (Wenneras and Wold, 1997). These concerns seem to be partly caused by the fact that the reviewer selection only includes a small sample from all peers potentially available. Aggravating this situation, no common agreements exist to provide reviewers with uniform guidelines, let alone binding rules, and no established standards by which those rules can be designed—peer review is in fact largely conducted at the discretion of the reviewers themselves. Given that reviewers vary considerably with respect to assessment and strictness, manuscript evaluation in the present model is highly dependent on reviewer selection. The lack of validity is further compounded by review and reviewer confidentiality, rendering them elusive to follow-up inspection.

Having been published, a scientific paper is exposed to interested scholars and hence goes through an ongoing process of *open evaluation*. When compared to journal-guided procedures, post-publication peer review is more suitable for evaluating

research impact, as scientists constantly need to consider which work they choose to accept, refute or expand upon. Over time, publications are thereby empirically detached from affixed quality labels like journal impact (but high-impact publications remain predominantly requested when it comes to promotions and grant applications, as journal publishing has traditionally been the main means of disseminating scientific knowledge). Although part of every individual scientist's everyday work, this communal effort has so far failed to develop into a cohesive framework within which research evaluation can be managed systematically and efficiently. A first step towards challenging this state was made feasible through the technological advancements of the web 2.0, constituting a change toward more openness between both researchers themselves, and researchers and public. This has manifested in the establishment of online open access formats and data repositories, and the growing recognition of scientific blogs and social networks for massive-scale scholarly exchange (e.g., <http://thirdreviewer.com>, <http://peerevaluation.org>). Such examples demonstrate the potential of exploiting the broad communication resources and simple usability of web-based technology by translating it into scientific practice. Smith (1999, 2003) reviews the current state of net-based publishing, concluding that all the activities of traditional journal publishing could be carried out collaboratively by existing web services. In a similar vein, overlay journals utilize the web to compile distributed information about one particular topic (Enger, 2005; Harnad, 2006). These studies indicate the high potential of distributed networking for a framework of open evaluation.

The principle of exchanging evaluation content through a data format and protocol has been put forward by Rodriguez et al. (2006). However, important elements of a framework, such as topic ("subject domain") attribution, review evaluation, reviewer selection, and other aspects central to the evaluation process, are in that system based on recursively data-mining the references of a paper. Even more so, single review elements are not evaluated on



their own worth, instead they are weighted by the reviewers life-time “influence”. Riggs and Wilensky (2001) come one step closer in that their rating of reviewers is based on the agreement with other reviews, yet they also do not differ between single reviews by the same reviewer. We deem these aspects a part of the evaluation, and think they should therefore be done by peers, case-by-case.

In the following article, we suggest utilizing the advantages of web-based communication in order to implement a framework of post-publication peer review. First, we outline its requirements, standard and protocol, serving to unify services dedicated to the evaluation of published research, and pinpoint its potential to help overcome shortcomings of the current reviewing system. We particularly emphasize the importance of provider-independence, with potentially infinite implementations. Finally, we illustrate our approach with a minimal working example and discuss it further.

## REQUIREMENTS FOR A NETWORK DEDICATED TO OPEN EVALUATION

Based on the weaknesses of pre-publication assessment, we formulate six criteria, which any net-based design aiming to attain large-scale open evaluation should be bound to fulfill in order to usefully complement the current state:

### ACCESSIBILITY FOR ALL AND EACH REGISTERED USER IS ENTITLED TO REVIEW

Open evaluation content should be open to everyone with an interest in science. Just as with pre-publication review, it builds on the expertise of peers; however, peer review does not need to be limited by external reviewer selection. Quite to the contrary, a network of open evaluation should recognize every user as a potential reviewer in order to most effectively serve to amass criticism. This should include scholars from topic-related and -unrelated fields as well as the educated public.

### EACH PUBLICATION CAN BE SUBJECT TO AN INFINITE NUMBER OF REVIEWS

A single review only represents a single opinion. At best, it was carried out objectively, identifies all flaws in a manuscript and contains helpful suggestions for improvement; even in this ideal case, standards between reviewers vary. At worst, a reviewer conducts reviewing according to career interests. The continuum between these two extremes is vast and impossible to ignore. However, objectivity can arguably be enhanced by incorporating many opinions. Hence, the number of reviews pertaining to a given publication needs to be unrestrained. If a large number of reviews conform to a particular opinion, it is likely that their assessment deserves notice. Even more importantly, if the dissimilarity between reviews is high, this indicates the need for further feedback from competent peers. Separate reviews can finally be consolidated into one complete assessment whose outcome reliably approximates the actual value of a publication.

### EACH REVIEW NEEDS TO BE DISCLOSED

The value of a scientific study depends on its recognition by other scientists. Careful feedback from the community is indispensably valuable for both the executing scientist and the recipient, as they

help to put results in perspective and can motivate adjustments or new research. In a network of open science evaluation, a review should be understood as just another type of publication directed to a topic-interested audience, including the author. Therefore, its disclosure is necessary in order to gauge its reception among other users, which will determine the overall quality. Since that way it is more likely to be scrutinized, it also serves as an incentive for thorough reviewing.

### REVIEWS NEED TO BE ASSESSABLE

Each review has to be considered potentially subjective, incomplete or faulty; consequently, it needs to be the subject of evaluation, just as with scientific publications. Here, a review of a review is termed *meta review*. Meta reviews can be either quantitative or qualitative (see “Standard”). Thereby, existing reviews can be rated and sorted by their overall reception. It further reduces repetitiveness and prevents trolling (i.e., posting off-topic comments). As per definition, any given meta review can again be target of another meta review, and so on.

Note that our design advocates information *gathering* rather than *re-computation*: after publication, an article has usually already received some level of attention; evaluation is carried out by individuals and journal clubs, lab meetings and other events devoted to research. Existing assessments thus often only need to be collected, and can then be analyzed and shared.

### REVIEWER EXPERTISE NEEDS TO BE DIFFERENTIATED THROUGH COMMUNITY JUDGMENT

For each user, expertise is bound to specific entities, such as a scientific method or theory, and among those, pronounced to varying degrees. User accounts thus need to feature a discernible expertise profile. An expertise profile should reflect scientific topics that were addressed by the user in submitted reviews and own publications. In turn, the attribution of these topics should be performed by peers.

### MANDATORY USER AUTHENTICATION

To ascertain the level of participation, account should be taken of the user’s authentication. Reading and submission of reviews should be enabled upon registration, where the user authenticates themselves with an unidentifiable credential such as a valid email address. Additionally, users may find it worthwhile to indicate their academic status with a validated academic email address, branding the account as “validated scientist” which may initially increase their perceived trustworthiness. Note that in general, authentication does not imply general disclosure of identity. In fact, authentication is necessary to unambiguously attribute the content to its real author, regardless of the level of anonymity. Nonetheless, it is conceivable that a user might prefer to disclose his identity (McNutt et al., 1990; Justice et al., 1998; Godlee, 2002; Bachmann, 2011), as it may add further credibility and acknowledgment to their effort. In order to illustrate and stimulate this initiative we formulate the following tentative initial features.

## EXCHANGING OPEN EVALUATION THROUGH AN IMPLEMENTATION-INDEPENDENT FRAMEWORK

The above described requirements initially invite the idea of an implementation as one platform. However, when comparing to



the current situation of both pre- and post-publication evaluation and distribution, such a platform seems economically unviable. Multiple institutions and companies compete for a role in these processes, so that their united participation is unlikely. In addition, dependence on one such system might be incongruous with the scientific principle of independent research. Therefore, we suggest that the above described structure should be implemented through a framework of: (1) a standard for the structure of the evaluation data and (2) a protocol for their communication, tentatively called *Framework for Open Science Evaluation* or FOSE (Figure 1). Both protocol and standard should be in the public domain. Ideally, such a framework should in time be agreed upon by all relevant parties. These include essentially the same organizations that may implement FOSE, such as academic institutions, publishers, funding agencies, scientist interest groups, etc. Supporting resources, such as open source software libraries, that implement the framework with an *Application Programming Interface (API)* and descriptive documentation could further promote the usage. This approach then supports the development of concrete platforms that make use of this framework, enabling them to share and integrate the evaluation content.

### RESPONSIBILITY

To develop and maintain this framework, an organization should be in place. This organization could be modeled after the successful W3C (<http://www.w3.org/>) organization which is responsible for the arguably daunting task of agreeing upon standards and protocols for the World Wide Web. Representatives of groups and organizations with an interest should be invited to participate, as their use and compliance is important to the success of this approach. Such members could include publishers, major research institutions, funding agencies, and the scientific community in general.

### STANDARD

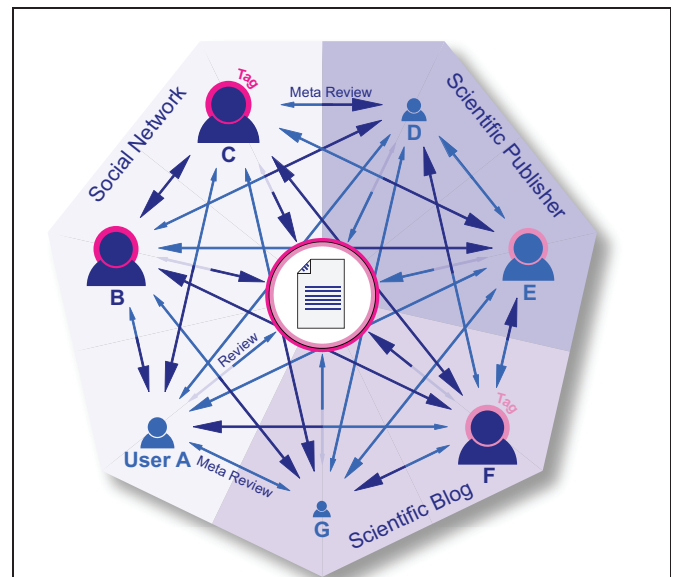
The framework structures the review process with *evaluation elements*. These predefined contributions can themselves contain standardized attributes.

#### Qualitative review

In the context of the proposed design, we define a qualitative review as any textual feedback of undefined format (length, level of detail). The content need not necessarily be of the appraising kind—questions with the goal of clarification and journal-club-style summaries are just some of the alternative content types that are at place here. Potential attributes may include a *creation timestamp* indicating the date and time it was submitted.

#### Quantitative review/rating

Content should also be evaluable with ordinal ratings. This way it can be quantitatively compared and sorted. Potential attributes include *contextualizing phrase*, which is a “category of feedback” indicating what aspect of the content the rating applies to, or *scale type*, denoting the range and order of the used scale as compared to a general reference scale.



**FIGURE 1 | Open evaluation can be organized through a uniform framework of open scientific social networking.** The figure explains the concept of an implementation-independent framework in the context of open evaluation. Importantly, reviewers can fulfill their roles via independent providers that all conform to the standard of the framework; an implementation as a single social network website is rendered unnecessary herewith. Three sectors of the pie chart correspond to three platforms implementing FOSE: a social network (either specifically academic or for other purposes), a scientific publisher, and a scientific blog. User icons labelled A to G represent participating reviewers. Arrows indicate their interactions with an original publication placed at the center (that is, submission of a *review*) and each other's evaluative content (submission of a *meta review*). Arrows coming from the center (semi-transparent) symbolize responses from the publication's author, who can participate with evaluation content. Light and dark pink halos represent thematic tags that have previously been assigned to both the publication and some users, namely reviewer B and C (dark pink) as well as E and F (light pink). Note that for simplicity, both tags have an equally high load. Icon size encodes authority (that is, a continuous variable indicating how proficiently a user has reviewed in the past, as seen by the community): larger size indicates higher authority, here exemplarily for a) tags shared with the publication in question (shown for users B, C, E and F), or b) other, publication-irrelevant tags (users A, D, G; not shown); it is important to stress that with respect to a given publication, reviews by users with either non-matching or no tags receive equal weights, irrelevant of their level of authority for other tags. The authority level for tags controls the weight and visibility of a user's review; review impact is here reflected by the size of the arrow heads, with a larger arrow head carrying more weight. Users B, C, and F already reached a critical authority threshold and were awarded an expert badge for their particular tag (see “Determining expertise and content classification” in Discussion); their nodes and arrows are therefore colored in dark blue. Since the publication has the same tags, their reviews have higher impact than those by other users. Note that E has indeed a tag but still lacks the necessary amount of positive ratings to appear as an expert. A, D, and G do not share any tags with the publication: their arrow heads are thus equally sized, indicating that they are not considered proficient in that scientific field.

#### Tag

A tag element attributes a certain topic to a target, that is, to a publication or to other evaluation content. Tags could either be retrieved from publications or proposed by a reviewer. A key

attribute is *tag load*, a continuous variable reflecting the number of reviewers that agree that the target covers a given topic. A tag can also be the target of a review element.

One critical attribute present in all elements is the *user identity* that is the unique and abstract reference to the user who has authored the content. Another important attribute is the *target*, referring to the content that is evaluated with a given element. In case of a regular review it would refer to a publication, whereas meta reviews target existing evaluation elements.

## PROTOCOL

In order to link elements of the evaluation such as those described above, the protocol must be able to unambiguously refer to publications. More importantly, the same should apply to connecting the elements among each other. This referring can be done through identifiers that are globally unique. Moreover, such elements pertaining to a limited set, for example all those applying to a certain publication and its evaluation elements, should be available for discovery. This requires a fixed address structure. Additional rules should apply to the referencing to the author of these elements. For anonymous elements, the reference should allow getting enough information to gauge authority (see “Determining Expertise and Content Classification” in the Discussion). For named elements, the reference should ideally be human-readable. This ensures ease of carrying over or referencing to their own identity by users.

Formula or rules could dictate how to gauge meta-level measures such as authority and impact, preferable on a content-dependent basis, as categorized by tags. As an example, the average user rating could be weighted amongst other indicators of quality (e.g., number of citations, the reviewee’s ratio of highly ranked publications, and other estimates) and appear in global score rankings. If the network is supervised by the community alone, review rating also counteracts malpractice. The protocol should further allow some customization by its implementing platforms, for example through additional extra-standard elements or attributes thereof.

## TECHNICAL IMPLEMENTATION

Earlier work by Rodriguez et al. (2006) proposed to integrate the evaluation content with the existing OAI-PMH framework for exchange of publication metadata. However, as Smith (1999, 2003) suggests, these are separate parts of the scientific process and, therefore, need not be served by the same system; indeed, this may be deleterious to their independent development. As with the current publisher-organized system for reviewing, OAI-PMH has centralized elements, and they propose to use one authoritative provider (Rodriguez et al., 2006, line 155, “The [...] pre-prints.”) Contrarily, in FOSE, there is no inherent difference between a provider and a consumer. However, for harvesting publications related to evaluative content, OAI-PMH would be a prime candidate. These requirements can, however, be partially implemented by the use of existing standards or technologies.

## IDENTIFIERS

A standard scheme for providing evaluation elements with an address could be based on the representational state transfer

### Example 1 | Localizing

A specific review hosted by provider Smith and Jones:

`smith-and-jones.com/fose/g3H2Ah4j`

All reviews of the paper with doi 888.444 hosted by provider Eval.net:

`eval.net/fose/reviews/by-doi/888.444/`

(REST) resource identifier (Fielding, 2000). That is, the address to the evaluation element (the resource) could be a unified resource locator (URL) formed from the provider’s address and several pre-defined hierarchical elements. Discovery would use the same scheme (see **Example 1**). Publications could be identified by means of the widely used Digital Object Identifier (DOI, e.g., Rosenblatt, 1997), and implementations could likely benefit from using a service such as OAI-PMH for the discovery of publications to evaluate.

In one approach, framework elements themselves can be created with a locally (at the host provider) unique identifier, such as a simple integer number key, or automatically generated random character sequence (e.g., UUID5, <http://tools.ietf.org/html/rfc4122>). They can then be referred to externally (at another provider or in another element) through the standardized address scheme. This has some drawbacks, most importantly, such a scheme is likely subject to “link rot,” if the implementation ceases to exist. The alternative, however, of centrally-registered links, such as the DOI, comes with a dependence on the registration agency, and in the case of a commercial agency, such as crossref, a financial cost (this would weigh in heavily when used for each review element), and does not support collections or discovery. These drawbacks seem to go against the idea of distributed responsibility, which is central to FOSE.

## DATA FORMAT

The documents can be encoded with XML (<http://www.w3.org/XML/>; see **Example 2**) and their format, as defined in the section

### Example 2 | Encoding

XML document excerpt using a FOSE namespace.

Qualitative review of publication with DOI 888.444:

```
<fose:qualitative-review id='smith-and-jones.com/fose/g3H2Ah4j'
created='03/05/2011' target-doi='888.444'
author='smith-and-jones.com/afarnsworth'>
```

Interesting manuscript. Please use gender-neutral pronouns.

```
</fose:review>
```

Quantitative meta review of the above review:

```
<fose:quantitative-review created='14/05/2011' target='smith-and-jones.com/fose/g3H2Ah4j'
author='eval.net/users/pbishop' scale=num>
8
</fose:review>
```

“Standard”, can be published and validated through the use of XML Schema (<http://www.w3.org/XML/Schema>).

Where (a) there is the need for the evaluation content to be machine readable (i.e., not just transferred, but “understood”), and (b) the structure of that content is complex enough not to be expressible in vanilla XML, it might be of advantage to publish that content in RDF. A strict XML scheme will allow RDF conversion from the XML. Further developments of standardization of data formats describing scientific knowledge (for instance along the line of nano-publications, Mons and Velterop, 2009) will naturally increase machine readability.

### A MINIMAL WORKING EXAMPLE

For the sake of simplicity, we illustrate the basic concept with an example of minimal complexity, involving one author (W. Bell), two scientist reviewers (O. Dunham and P. Bishop), and three independent services implementing FOSE (*The Journal of Foomatics*, the Medical Research Institution (NIX) employee site, *Eval.net*; Figure 2).

#### THE PUBLISHING AUTHOR

W. Bell’s research article “Effects of Cortexiphan on Inter-dimensional Travel” has successfully passed the pre-publication review and is now published in *The Journal of Foomatics*. A week later, Bell finds a notification of a new review in his inbox. This service is offered by *The Journal of Foomatics*, whose implementation of FOSE allows them to track reviews of their publications. In response to the review, he uses the *Journal of Foomatics* website

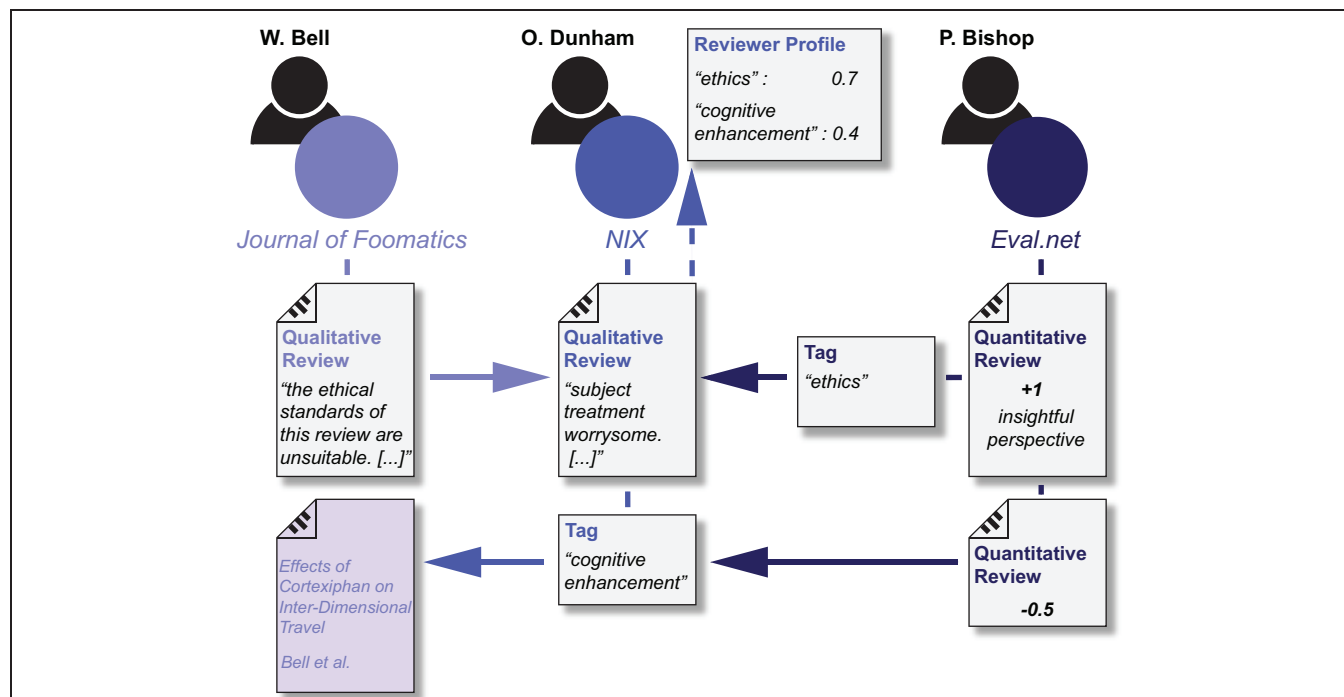
to comment on the referenced, unreasonably harsh, ethical standards. He further notices that another reviewer has supported the review’s assessment by giving it a positive rating.

#### THE REVIEWER

O. Dunham, a scientist working for the NIX, comes across Bell’s article during her literature research and, after reading, decides to publish some critical remarks about its ethical standards. She logs in to the NIX employee website and selects the tab “Review” where she finds textual and numerical rating elements; these features accord to the FOSE standard. Dunham submits a brief textual review on the ethical aspects of the subject’s involvement in the study. Moreover, she gives Bell’s publication the tag “cognitive enhancement” which starts out with an initial value of +1. Some days later, Dunham logs in to the NIX website and finds that a reviewer recently uprated her review on Bell’s article and assigned a tag “ethics”. Thereby, the *authority rating* of her reviewer profile (see “Assessing Scientific Impact” in Discussion) in “ethics” increases, and less so for “cognitive enhancement”, as this tag was downrated by Bishop. Moreover, this renders all of her reviews in those fields more visible to the community.

#### THE META REVIEWER

A PhD student named P. Bishop logs in to a community website called *Eval.net*, which has been designed by a non-profit organization with the aim of facilitating the scientific discourse. Again, *Eval.net* subscribes to FOSE and has downloaded



**FIGURE 2 | A minimal working example of open evaluation in FOSE.**

Depiction of example interactions of evaluation contributors concerning one publication, with an author (W. Bell), a reviewer (O. Dunham), and a meta

reviewer (P. Bishop). All are registered users of independent providers.

Arrows indicate targets of the evaluation content pieces they contribute. See "A Minimal Working Example" for complete narrative.

Dunham's review from *NIX*. Being interested in related topics, Bishop comes across Bell's article and Dunham's review. After having read both, he thinks that Dunham's review is valuable and rates it +1, *insightful perspective*. He, however, disagrees with the tag "*cognitive enhancement*" and gives it a quantitative review -0.5. He further assigns Dunham's review the tag "*ethics*."

## DISCUSSION

### ONE FRAMEWORK FOR EXCHANGE OF RESEARCH EVALUATION

Open evaluation, by its very nature, is a diverse approach: it aims at sourcing article evaluations in large quantities in order to approximate the value of a publication. In the web 2.0, it appears counter-intuitive to bind this process to a single community website. In fact, users should be free to choose from a range of independent providers with services tailored to specific interests and needs (users differ in interests and thus frequent different websites). For instance, a university's personnel platform (See *NIX website* in *A minimal working example*) may offer single sign-on for their employees, or a publisher has immediate access to publications. We believe that provider-variety significantly increases the overall participation of scientists and non-scientists in peer review. However, the main holdback for current platforms might be their closed, egocentric approach, which due to commercial motivations will not be readily accepted by other, influential contenders, thereby scattering the content. Instead of competing with other platforms, a new approach should promote interoperability. As a consequence, they must subscribe to one established norm in order to integrate evaluation content between them. The attractiveness of the framework for potential implementers should then be access to other, existing content at other providers. This integration ensures that all evaluation content is accessible everywhere, enhancing its traceability and comparability (across borders of papers, publishers, providers).

### MASSIVE AND IMMEDIATE FEEDBACK TO THE AUTHOR, THE SCIENTIFIC COMMUNITY, AND THE JOURNALS

Two shortcomings of the journal-based system have been widely mentioned: first, it obscures the discussion between authors and reviewers; second, there are few possibilities to comment on the result (Wicherts et al., 2012). Open evaluation can attenuate these weaknesses by administering feedback from the community to authors, reviewers, and publishers. We think that the proposed framework can help organize this process in a principled way. Feedback can be submitted online on institutional websites and platforms of third-party suppliers, and then collectively analyzed. *Ad-hoc* networking will enable users—authors and reviewers, scientists and the educated public—to engage in discussions among themselves. In combination, these features craft a highly transparent and dynamic alternative to established means of article evaluation, such as response letters and scientific meetings, and have several advantages over them: first, authors will receive unfiltered criticism by the scientific community in a quantitative and qualitative manner. The higher the participation, the more meaningfully does the evaluation in sum approximate the benefit for the community. To reflect that, numerical ratings could be related to the number of submitted reviews. However,

even few reviews are likely to contain valuable feedback, considered that they come from a vast pool of potential reviewers. Second, any feedback given is instantly visible to the scientific community and can thus be challenged and questioned. As a consequence, reviews can be commented and rated in turn. Third, the network generally encourages discussions about scientific publications. An additional advantage of post-publication review is feedback for journals. The ratings obtained in open evaluation can be compared to the editor's assessment used for the publication decision. If there is a large discrepancy, the journal could change its assessment policy or reviewer selection and instruction.

### DETERMINING EXPERTISE AND CONTENT CLASSIFICATION

FOSE sources reviews as globally as possible, setting few constraints on general participation. This raises the issue of trustworthiness, as reviewers clearly differ in their authority with respect to a particular scientific field. Drawing a bold line between affiliated and unaffiliated scholars (as by email authentication only) is insufficient to resolve this and would allay criticism from the latter. In fact, it is equally necessary to distinguish between experts and laymen within a given scientific field. Therefore, in a framework with no prior reviewer selection, expertise is only determinable *post-hoc*. Natural sources for this assessment are reviews and meta reviews. Someone should arguably be considered proficient in a specific theory, method, research area, etc., when he or she has garnered a critical mass of positive evaluation on publications, i.e., articles and reviews. By reaching a certain threshold, a user could be awarded an "expert" badge. This would provide a communally determined credential distinguishing proficient users from others. Consequently, their reviews should carry greater weight and be most visible; they could be branded *expert review* and analyzed both separately and jointly with non-expert reviews, with scores displayed in the user statistics. Complementary to that binary classification, expertise should also be recognized as a continuum. The level at which any user is authorized to contribute could generally refer to a variable named *authority*: again, one's authority should depend on the average quality of one's reviews submitted, and reviews by users with lower *authority* should be less visible (**Figure 1**). One critical ingredient in this formula is content classification. To restrict *authority* to a given field, contributions must be classified as covering a particular theme. In the proposed framework, peer-based tagging fills this role. Thereby, the community can discuss their opinion on the attribution of these tags through quantitative and qualitative reviews. The use of tags leaves space for advanced indexing methods, such as hierarchical relations between tags, and is a further step in the direction of more advanced semantic markup, such as nano-publications (Mons and Velterop, 2009). One's individual reviewer profile could be determined by the union of all tags targeting the user's contributions, weighted by participation and content quality. This accrued information could also be used by automatic filters in order to suggest other publications for review. In a more general sense, by jointly crowd-sourcing scientific classification and evaluation, tags can be utilized to meaningfully and reliably index scientific literature.



## ASSESSING SCIENTIFIC IMPACT

An interoperable framework lends itself ideally to assessing the scientific impact of a publication. An article that is highly relevant to many scientists will attract more attention and will receive more and, on average, higher ratings than one being less so. Quantitative reviews in the form of numerical ratings can be summed up in statistics linked to user profiles. Statistics could feature different components, such as the average rating given to all publications, separate averaged ratings for reviews by scientists and the general public as well as their union, average meta review ratings, evaluation of replicability, etc. It will be one major assignment of the development of a universal standard to define meaningful scales for quantitative reviews. Contrary to pre-publication review, this will bring about an assessment model of scientific literature, in which evaluation is sought from *many* people and within a technically infinite period. Hence, it remains amendable over time: statistics of a publication can always be up- or downgraded by a new review. They are thereby more likely to reflect the actual scientific impact of a publication on the scientific endeavour. This approach makes a contribution toward counterbalancing the current focus of science on journal prestige, contrary to which it cooperatively approximates the value of a scientific publication based on actual relevance. It is desirable that these advantages are also accounted for in practical ramifications associated with publications. To that end, we believe that post-publication reviews can just as well serve as a valid reference in hiring processes and grant applications as publications in prestigious journals. As they provide an independent quality indicator for each publication, they should be referred to in an author's quantitative evaluation and used to put journal prestige in perspective (e.g., when an article in a low-impact journal receives a lot of attention or vice versa). Similarly, positively recognized reviewing can provide a reference on a user's scientific expertise even in the absence of own publications. The possibility to refer to one's user statistics as a meaningful reference in turn will incentivize participation in such a framework: influential research or expertise in a particular field is likely to be recognized by the community, even more so over time. One's reviewer profile could be added to one's CV in order to distinguish oneself. In the same vein, authors of scientific publications should be enabled to refer to their reception in order to add another plus to their resume. Moreover, as it reduces attention to journal impact,

it will adjust the allocation of scarce resources (i.e., positions, grants, etc.) on the basis of scientific soundness; hence an excellent article will be more likely to receive credit, regardless where it was published.

## OPENING THE SCIENTIFIC DEBATE

Scientific progress critically depends on the interaction among scientists. Traditionally, the dissemination of scientific content is achieved mainly by two means: scientific conferences, closed to the outside and limited to a usually pre-selected group of scientists; and publications, whose review process is only visible to a handful of people. Furthermore, only few scientific findings are chosen to be translated to the larger public (which at that point have been subjected to massive informational filtering and simplification). Hence, from society's perspective, the production of new knowledge can hardly be seen as participative or fully transparent. This status quo is inadequate, as scientific work arguably depends on society's endorsement, which requires mutual understanding, hence transparency. In that vein, the boundaries between science and the public domain have recently been blurred by the emergence of net-based communication (see "Introduction"). FOSE contributes to this development in that it utilizes social networking in order to help transcend the barrier between science and general public by open user-to-user propagation of scientific knowledge. Even more so, it is able to integrate criticism of scientific papers from *non-scientist* reviewers, yielding a more complete picture of research evaluation.

## TOWARD OPEN EVALUATION OF SCIENCE

The organization of research evaluation will always be competed for by a multitude of players; integrating their contributions into a cohesive framework promises the most efficient way to aggregate peer review, and ultimately to reliably reflect scientific impact. Accordingly, the FOSE way to open evaluation is *open*, that is by exchange between providers, and through *evaluation*, that is by having peers recursively evaluate content. These two principles rest on a standard for structuring this content and a protocol for its exchange.

## ACKNOWLEDGMENTS

Both authors would like to thank their host institutes for funding their work.

## REFERENCES

- Bachmann, T. (2011). Fair and open evaluation may call for temporarily hidden authorship, caution when counting the votes, and transparency of the full pre-publication procedure. *Front. Comput. Neurosci.* 5:61. doi: 10.3389/fncom.2011.00061
- Casati, E., Marchese, M., Mirylenka, K., and Ragone, A. (2010). Reviewing peer review: a quantitative analysis of peer review. Technical Report DISI-10-014, *Ingegneria e Scienza dell'Informazione*.
- Enger, M. (2005). *The Concept of 'Overlay' in Relation to the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)*. Master's thesis, Universitet i Tromsø, Tromsø, Norway.
- Fielding, R. T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. Doctoral dissertation, University of California, Irvine.
- Godlee, F. (2002). Making reviewers visible. *Am. Med. Assoc.* 287, 2762–2765.
- Godlee, F., Gale, C. R., and Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports a randomized controlled trial. *Am. Med. Assoc.* 280, 237–240.
- Goodman, S. N., Berlin, J., Fletcher, S. W., and Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at annals of internal medicine. *Ann. Intern. Med.* 121, 11–21.
- Harnad, S. (2006). Research journals are already just quality controllers and certifiers: so what are "overlay journals"? Available online at: <http://goo.gl/rzunU>
- Justice, A. C., Cho, M. K., Winker, M. A., Berlin, J. A., and Rennie, D. (1998). The PEER investigators. Does masking author identity improve peer review quality: a randomized controlled trial. *Am. Med. Assoc.* 280, 240–242.
- McNutt, R. A., Evans, A. T., Fletcher, R. H., and Fletcher, S. W. (1990). The effects of blinding on the quality of peer review. A randomized trial. *Am. Med. Assoc.* 263, 1371–1376.



- Mons, B., and Velterop, J. (2009). "Nano-publication in the e-science," in *Proceeding of Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, (Washington, DC).
- Peters, D. P., and Ceci, S. J. (1982). Peer-review practices of psychological journals: the fate of published articles, submitted again. *Behav. Brain. Sci.* 5, 187–195.
- Riggs, T., and Wilensky, R. (2001). "An algorithm for automated rating of reviewers," in *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL'01*, (New York, NY, USA: ACM), 381–387.
- Rodriguez, M. A., Bollen, J., and van de Sompel, H. (2006). The convergence of digital libraries and the peer-review process. *J. Inf. Sci.* 32, 149–159.
- Rosenblatt, B. (1997). The digital object identifier: solving the dilemma of copyright protection online. *J. Electron. Publ.* 3, 2.
- Smith, J. W. T. (1999). The deconstructed journal—a new model for academic publishing. *Learn. Publ.* 12, 79–91.
- Smith, J. W. T. (2003). "The deconstructed journal revisited—a review of developments," in *From Information to Knowledge: Proceedings of the 7th ICC/IFIP International Conference on Electronic Publishing held at the Universidade do Minho*, eds S. M. de Souza Costa, J. A. A. Carvalho, A. A. Baptista, and A. C. S. Moreira, (Portugal: ELPUB).
- Wenneras, C., and Wold, A. (1997). Nepotism and sexism in peer-review. *Nature* 387, 341–343.
- Wicherts, J. M., Kievit, R. A., Bakker, M., and Borsboom, D. (2012). Letting the daylight in: reviewing the reviewers and other ways to maximize transparency in science. *Front. Comput. Neurosci.* 6:20. doi: 10.3389/fncom.2012.00020
- A. A. Baptista, and A. C. S. Moreira, (Portugal: ELPUB).
- that could be construed as a potential conflict of interest.

Received: 16 May 2011; accepted: 21 May 2012; published online: 27 June 2012.

Citation: Walther A and van den Bosch JFF (2012) FOSE: a framework for open science evaluation. *Front. Comput. Neurosci.* 6:32. doi: 10.3389/fncom.2012.00032

Copyright © 2012 Walther and van den Bosch. This is an open-access article distributed under the terms of the Creative Commons Attribution Non Commercial License, which permits non-commercial use, distribution, and reproduction in other forums, provided the original authors and source are credited.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships



# Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation

Ulrich Pöschl \*

Max Planck Institute for Chemistry, Mainz, Germany

## Edited by:

Diana Deca, Technical University  
Munich, Germany

## Reviewed by:

Alessandro Treves, Scuola  
Internazionale Superiore di Studi  
Avanzati, Italy  
Christian Leibold, Ludwig Maximilians  
University, Germany  
Erik Sandewall, Linköping University,  
Sweden

## \*Correspondence:

Ulrich Pöschl, Max Planck Institute for  
Chemistry, Hahn-Meitner-Weg 1,  
D-55128 Mainz, Germany.  
e-mail: u.poeschl@mpic.de

The traditional forms of scientific publishing and peer review do not live up to all demands of efficient communication and quality assurance in today's highly diverse and rapidly evolving world of science. They need to be advanced and complemented by interactive and transparent forms of review, publication, and discussion that are open to the scientific community and to the public. The advantages of open access, public peer review, and interactive discussion can be efficiently and flexibly combined with the strengths of traditional scientific peer review. Since 2001 the benefits and viability of this approach are clearly demonstrated by the highly successful interactive open access journal Atmospheric Chemistry and Physics (ACP, [www.atmos-chem-phys.net](http://www.atmos-chem-phys.net)) and a growing number of sister journals launched and operated by the European Geosciences Union (EGU, [www.egu.eu](http://www.egu.eu)) and the open access publisher Copernicus ([www.copernicus.org](http://www.copernicus.org)). The interactive open access journals are practicing an integrative multi-stage process of publication and peer review combined with interactive public discussion, which effectively resolves the dilemma between rapid scientific exchange and thorough quality assurance. Key features and achievements of this approach are: top quality and impact, efficient self-regulation and low rejection rates, high attractivity and rapid growth, low costs, and financial sustainability. In fact, ACP and the EGU interactive open access sister journals are by most if not all standards more successful than comparable scientific journals with traditional or alternative forms of peer review (editorial statistics, publication statistics, citation statistics, economic costs, and sustainability). The high efficiency and predictive validity of multi-stage open peer review have been confirmed in a series of dedicated studies by evaluation experts from the social sciences, and the same or similar concepts have recently also been adopted in other disciplines, including the life sciences and economics. Multi-stage open peer review can be flexibly adjusted to the needs and peculiarities of different scientific communities. Due to the flexibility and compatibility with traditional structures of scientific publishing and peer review, the multi-stage open peer review concept enables efficient evolution in scientific communication and quality assurance. It has the potential for swift replacement of hidden peer review as the standard of scientific quality assurance, and it provides a basis for open evaluation in science.

**Keywords:** open evaluation, public peer review, open access publishing, interactive discussion, open peer commentary, transparency, self-regulation

## INTRODUCTION

The traditional ways of scientific publishing and peer review do not live up to the needs of efficient communication and quality assurance in today's highly diverse and rapidly developing world of science. Besides high profile cases of scientific fraud, science, and society are facing a flood of carelessly prepared scientific papers that are locked away behind subscription barriers, dilute rather than enhance scientific knowledge, lead to a waste of resources and impede scientific and societal progress. On the other hand, the spread of innovative ideas and concepts is often delayed by inertia and obstruction in the hidden review process of traditional mainstream scientific journals (Pöschl, 2004).

Open access to scientific research publications is desirable for many educational, economic, and scientific reasons (Max Planck Society, 2003; David and Uhler, 2005; European Commission and German Commission for UNESCO, 2008), and it provides major opportunities for the improvement of scientific communication, quality assurance, and evaluation (Bodenschatz and Pöschl, 2008; Pöschl and Koop, 2008; Pöschl, 2010b):

- (1) Open access is fully compatible with traditional peer review, and in addition it enables interactive and transparent forms of review and discussion open to all interested members of the scientific community and the public (open peer review).

- (2) Open access gives reviewers more information to work with, i.e., it provides unlimited access to relevant publications across different scientific disciplines and communities (interdisciplinary scientific discussion and quality assurance).
- (3) Open access facilitates the development and implementation of new metrics for the impact and quality of scientific publications (combination of citation, download/usage, commenting, and ranking by various groups of readers and users, respectively; Bollen et al., 2009).
- (4) Open access helps to overcome the obsolete monopoly/oligopoly structures of scientific publishing and statistical analysis of publication contents and citations/references, which are limiting the opportunities for innovation in scientific publishing and evaluation.

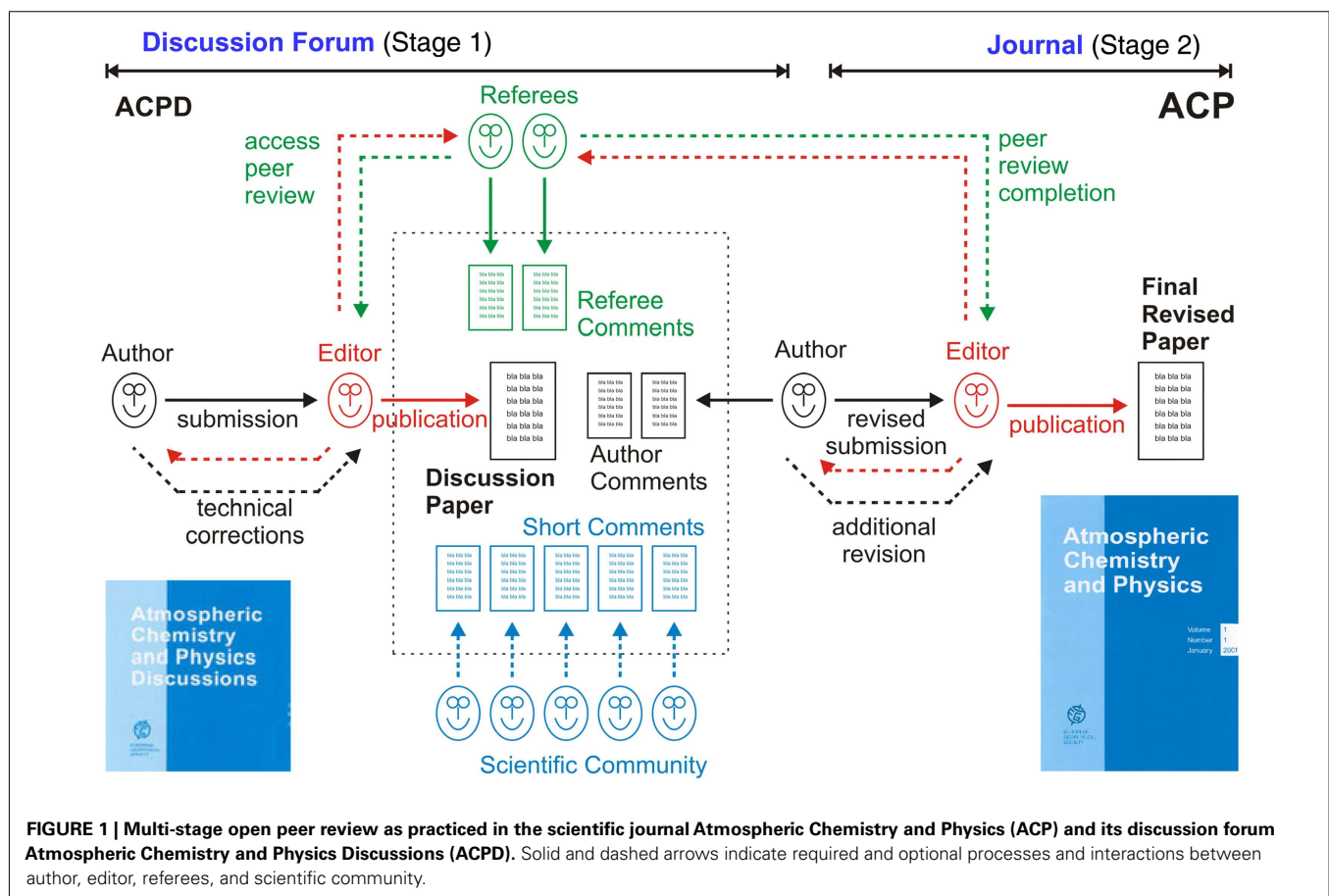
As demonstrated below, the effects and advantages of open access, public review, and interactive discussion can be efficiently and flexibly combined with the strengths of traditional scientific publishing and peer review (Pöschl, 2009a, 2010a,b). Unlike other, more radical proposals of how to change and improve scientific quality assurance, the interactive open access publishing approach introduced by the international scientific journal Atmospheric Chemistry and Physics (ACP) conserves the strengths of traditional peer review while overcoming its major weaknesses. This approach is compatible with the structures of traditional scientific publishing and quality assurance, and thus it enables

an efficient transition from the operational but sub-optimal past of subscription-based journals and hidden peer review to the future of free exchange and transparent evaluation of scientific information on the internet.

## MULTI-STAGE OPEN PEER REVIEW

So far, the arguably most successful alternative to the closed peer review of traditional scientific journals is the multi-stage open peer review practiced by ACP and a growing number of interactive open access sister journals of the European Geosciences Union (EGU) and Copernicus Publications (Pöschl, 2010b). As detailed below (see Atmospheric Chemistry and Physics and the European Geosciences Union), ACP is by most if not all standards more successful than comparable scientific journals with traditional or alternative forms of peer review (editorial statistics, publication statistics, citation statistics, economic costs, and sustainability). The multi-stage open peer review of ACP is based on a two-stage process of open access publishing combined with multiple steps of peer review and interactive public discussion as illustrated in **Figure 1**.

In the first stage, manuscripts that pass a rapid pre-screening (access review) are immediately published as “discussion papers” in the journal’s discussion forum (Atmospheric Chemistry and Physics Discussions, ACPD). They are then subject to interactive public discussion for a period of 8 weeks, during which the comments of designated referees, additional comments by other interested members of the scientific community, and the authors’



replies are also published alongside the discussion paper. While referees can choose to sign their comments or remain anonymous, comments by other scientists (registered readers) are automatically signed. In the second stage, manuscript revision and peer review are completed in the same way as in traditional journals (with further rounds of review and revision where required) and, if accepted, final papers are published in the main journal. To provide a lasting record of review and to secure the authors' publication precedence, every discussion paper, and interactive comment remains permanently archived and individually citable.

The multi-stage peer review and publication process of ACP effectively resolves the dilemma between rapid scientific exchange and thorough quality assurance, and it offers a win-win situation for all involved parties (authors, referees, editors, publishers, readers/scientific community). The primary positive effects and advantages compared to the traditional forms of publication with closed peer review are:

1. The discussion papers offer free speech and rapid dissemination of novel results and original opinions, without revisions that might delay or dilute innovation (authors' and readers' advantage).
2. The interactive peer review and public discussion offer direct feedback and public recognition for high quality papers (authors' advantage); they prevent or minimize the opportunity for hidden obstruction and plagiarism (authors' advantage); they provide complete and citable documentation of critical comments, controversial arguments, scientific flaws, and complementary information (referees' and readers' advantage); they reveal deficiencies and deter submissions of carelessly prepared manuscripts, thus helping to avoid/minimize the waste of time and effort for deficient submissions (referees', editors', publishers', and readers' advantage).
3. The final revised papers offer a maximum of scientific information density and quality assurance achieved by full peer review (with optional anonymity of referees) and revisions based on the referees' comments plus additional comments from other interested scientists (readers' advantage).

Readers who are primarily interested in the quintessence of manuscripts that have been fully peer reviewed and approved by referees and editors can simply focus on the final revised paper (or, indeed, its abstract) published in the journal and neglect the preceding discussion papers and interactive comments published in the discussion forum. Thus the two-stage publication process does not inflate the amount of time required to maintain an overview of final revised papers. On the other hand, readers who want to see original scientific manuscripts and messages before they are influenced by peer review and revision, and who want to follow the scientific discussion between authors, referees, and other interested scientists, can browse the papers and interactive comments in the discussion forum.

The possibility of comparing a final revised paper with the preceding discussion paper and following the interactive peer review and public discussion also facilitates the evaluation of individual publications for non-specialist readers and evaluators. The

style and quality of interactive commenting and argumentation provide insights that go beyond, and complement, the information contained in the research article itself.

The multi-stage process of review and publication stimulates scientists to prove their competence via individual high quality papers and their discussion, rather than just by pushing as many papers as possible through journals with closed peer review and no direct public feedback and recognition for their work. Authors have a much stronger incentive to maximize the quality of their manuscripts prior to submission for peer review and publication, since experimental weaknesses, erroneous interpretations, and relevant but unreferenced earlier studies are more likely to be detected and pointed out in the course of interactive peer review and discussion open to the public and all colleagues with related research interests.

Moreover, the transparent review process prevents authors from abusing the peer review process by delegating some of their own tasks and responsibilities to the referees during review and revision behind the scenes. Referees often make substantial contributions to the quality of scientific papers, but in traditional closed peer review their input rarely receives public recognition. The full credit for the quality of a paper published in a traditional journal generally goes to the authors, even when they have submitted a carelessly prepared manuscript that has taken a lot of time and effort on the part of the referees, editors, and publishers to turn it into a good one. While peer review depends crucially on the availability and performance of referees, it has traditionally offered little reward for those providing careful and constructive reviews. In public review, however, referees' arguments are publicly heard and, if comments are openly signed, referees can also claim authorship for their contribution.

Note that most of the effects and advantages outlined above are not fully captured by alternative approaches where interactive commenting and public discussion occur only after formal peer review and final publication of scientific papers or where the discussion paper and interactive comments are removed after publication of the final revised paper (see Key features of multi-stage open peer review as practiced by ACP).

Overall, the interactive open access publishing philosophy emphasizes the value of free speech and efficient public exchange and scrutiny of scientific results in line with the principles of critical rationalism and open societies. Accordingly, editors and referees are supposed to critically comment and evaluate manuscripts, to help authors improve their manuscripts, and to eliminate clearly deficient manuscripts. However, authors shall not be forced to adopt the editors' or referees' views and preferences. Instead, the readers shall be able to make up their own mind in view of the public review and discussion. In case of doubt, editorial decisions shall favor free speech of scientists, and in the end, scientific progress; history shall tell if – or to which degree – they were right. In scientific research, the line between fundamental flaws and major innovations can be fine, and the multi-stage process of interactive open access publishing and peer review enables efficient balancing and differentiation between potentially misleading hypotheses and innovative theories even in highly controversial cases (Pöschl, 2004, 2010b).

## ATMOSPHERIC CHEMISTRY AND PHYSICS AND THE EUROPEAN GEOSCIENCES UNION

The interactive open access journal Atmospheric Chemistry and Physics (ACP<sup>1</sup>), founded in 2001, demonstrates that multi-stage open peer review enables much more efficient quality assurance than traditional closed peer review. ACP is run by the European Geosciences Union (EGU<sup>2</sup>), the open access publisher Copernicus<sup>3</sup>, and a globally distributed network of scientists (~130 co-editors coordinated by an executive committee of five). Manuscripts are normally handled by an editor who is familiar with the specific subject area of the submitted work and independently guides the review process. Details about the largely automated handling and editor assignment of submitted manuscripts are given below (see Key features of multi-stage open peer review as practiced by ACP) and on the journal website. The origin and development of interactive open access publishing as practiced by ACP and EGU/Copernicus are specified in a recent anniversary publication (Pöschl, 2010c, 2011; Copernicus, 2011)<sup>4</sup>.

Currently ACP publishes about 800 papers per year (~13,000 double column print pages), which is similar to the volume of traditional major journals in the fields of chemistry and physics (ISI Science Citation Index, Journal Citation Report, 2010). On average, each paper receives four interactive comments, and about one in five papers receives a comment from the scientific community in addition to the comments from designated referees. In total, there are typically 0.5 pages of interactive comments per page of original discussion paper, i.e., the volume of interactive comments amount to as much as ~50% of the volume of discussion papers. The interactive comments show the full spectrum of opinions in the scientific community, ranging from harsh criticism to open applause (sometimes for the same discussion paper), and they provide a wealth of additional information and evaluation that is available to everyone.

About three out of four referee comments are posted without the referee's name, showing that most referees in the scientific community of ACP prefer anonymity. There are, however, interesting differences between sub-disciplines: on average about 20% of theoreticians and computer modelers sign their referee comments, while only 10% of the laboratory and field experimentalists do so. It appears that modelers more often provide suggestions and ideas for which they like to claim authorship as a reward. The anonymous referee comments are generally also very constructive and substantial. The ACP editors do not actively moderate the public discussions but reserve the right to delete abusive or inappropriately worded comments. Out of the nearly 20,000 interactive comments that have been posted so far, only a handful were removed or replaced because of inappropriate wording, which demonstrates efficient self-regulation by transparency.

Some colleagues have expressed concerns that referees may lose their independence by having access to the comments from fellow referees and from the public. Indeed, referees with limited

capacities occasionally seem to duplicate or refer to earlier comments without making up their own mind, but this is fairly easy to recognize and to take into account by editors and readers. Much more often, however, referees constructively build on or contradict earlier comments, which enhances the efficiency of review and discussion substantially. In theory, the independence of referees could be maintained by keeping submitted referee comments non-public until all referees have submitted their comments and these are all together published at the same time. In practice, however, this would cause unnecessary delays ("waiting for the last referee") and stifle rather than promote interactive discussion. Overall, experience shows that the advantages of enabling direct interaction between referees clearly outweigh the disadvantages.

The average rate of public commenting in addition to the designated referees' and authors' comments specified above (~20%) may appear low at first sight. It is, however, by an order of magnitude (factor ~10) higher than in journals with post-peer review online commenting and in traditional journals without online commenting (about 1–2%; Müller, 2008; Pöschl and Koop, 2008; Pöschl, 2010b). Discussion papers reporting controversial findings or innovations attract many interactive comments (up to 30 and more, see "Most commented papers" in the ACPD online library<sup>5</sup>). As expected, non-controversial papers usually elicit comments only from the designated referees. Why would scientists invest effort and time commenting on papers which they find interesting but not controversial?

In most scientific disciplines and journals (certainly in the fields of physics, chemistry, and biology with which the author is well acquainted) it is notoriously difficult to assign a couple of competent referees to every manuscript submitted for publication. In fact, this is the main bottleneck of peer review and scientific quality assurance, and most journal editors have to apply lots of manpower and electronic tools (invitation and reminder emails, etc.) to obtain a couple of referee comments per manuscript. Accordingly, the initiators and editors of ACP are quite satisfied with the overall number and volume of interactive comments. Higher rates of commenting were not expected and are not required to stimulate self-regulation mechanisms of scientific quality assurance (Pöschl, 2004, 2010a,b).

The editorial and citation statistics of ACP clearly demonstrate that multi-stage open peer review indeed facilitates and enhances scientific communication and quality assurance. The journal has relatively low rejection rates (~15% as opposed to ~50% in comparable traditional journals, Schultz, 2010), but only a few years after its launch ACP had already achieved top reputation and visibility in the scientific community. Accordingly, it quickly reached and maintained one of the highest ISI impact factors of several 100 journals indexed across the disciplines of atmospheric sciences, geosciences, and environmental sciences (JIF ≈ 5). These figures clearly confirm that anticipation of public peer review and discussion deters authors from submitting low-quality manuscripts and, thus, relieves editors and referees from spending too much time on deficient submissions. This is particularly important, because refereeing capacities are the most limited resource in scientific publishing and quality assurance. The high efficiency, robustness,

<sup>1</sup> [www.atmos-chem-phys.net](http://www.atmos-chem-phys.net)

<sup>2</sup> [www.egu.eu](http://www.egu.eu)

<sup>3</sup> [www.copernicus.org](http://www.copernicus.org)

<sup>4</sup> [http://www.atmospheric-chemistry-and-physics.net/general\\_information/public\\_relations.html](http://www.atmospheric-chemistry-and-physics.net/general_information/public_relations.html)

<sup>5</sup> [http://www.atmos-chem-phys-discuss.net/most\\_commented\\_papers.html](http://www.atmos-chem-phys-discuss.net/most_commented_papers.html)



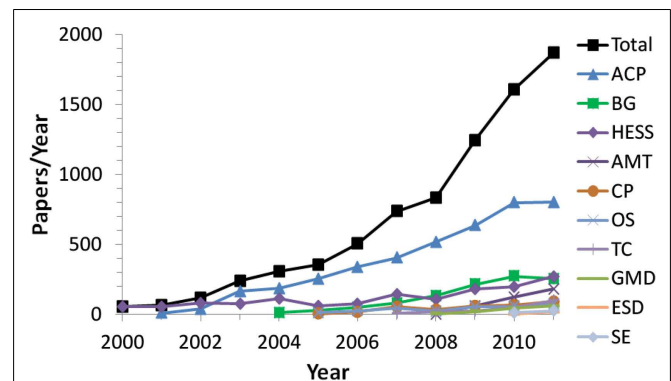
and predictive validity of the multi-stage open peer review process of ACP have been confirmed in a series of dedicated studies by evaluation experts from the social sciences (Bornmann and Daniel, 2010a,b; Bornmann et al., 2010, 2011a,b).

Since its launch in 2001, the number of articles published in ACP has increased rapidly. The high and increasing rates of submission, publication, and citation show that the scientific community values the open access, high quality, and interactive discussions of ACP. They confirm that there is a demand for improved scientific publishing and quality assurance, and that the interactive open access journal concept of ACP meets this demand. Today ACP is the largest journal in the field of atmospheric sciences and one of the largest across the fields of environmental and geosciences, offering at the same time top visibility and low rejection rates (2/5 year impact factors 5.4/5.8, rejection rate 15%, 12,000 pages in 2010). The combination of top visibility with high volume and low rejection rate, i.e., high efficiency by self-regulation, is a fairly unique achievement in the world of scientific publishing, where the most visible journals traditionally had relatively small volumes and high rejection rates (Copernicus, 2011; Pöschl, 2011).

Following up on the successful development of ACP, the EGU, and Copernicus have launched and are operating over a dozen of interactive open access sister journals in the geosciences and related disciplines, and more are in the pipeline<sup>6</sup>:

- Atmospheric Chemistry and Physics (ACP)<sup>7</sup>,
- Atmospheric Measurement Techniques (AMT)<sup>8</sup>,
- Biogeosciences (BG)<sup>9</sup>,
- Climate of the Past (CP)<sup>10</sup>,
- Drinking Water Engineering and Science (DWES)<sup>11</sup>,
- Earth System Dynamics (ESD)<sup>12</sup>,
- Earth System Science Data (ESSD)<sup>13</sup>,
- Geoscientific Instrumentation, Methods and Data Systems (GI, geoscientific-instrumentation-methods-and-data-systems.net),
- Geoscientific Model Development (GMD)<sup>14</sup>,
- Hydrology and Earth System Sciences (HESS)<sup>15</sup>,
- Ocean Science (OS)<sup>16</sup>,
- Social Geography (SG)<sup>17</sup>,
- Solid Earth (SE)<sup>18</sup>,
- The Cryosphere (TC)<sup>19</sup>.

**Figure 2** illustrates the growth of ACP and the other EGU interactive open access journals over the past decade<sup>20</sup>. The wide range



**FIGURE 2 |** Number of papers published per year in the interactive open access journals of the European Geosciences Union (EGU).

of different topics and scientific communities covered by the EGU interactive open access journals demonstrates that multi-stage open peer review is suitable for any kind of topical scientific journal. For example, the community of cryospheric sciences is much smaller than that of atmospheric sciences, but the development of the cryospheric science journal (TC) proceeds at least as well as that of the atmospheric science journals (ACP and AMT). The first journal impact factor of TC was already the highest in its field. The journal Hydrology and Earth System Sciences (HESS) had already existed as a subscription-based journal with traditional peer review before it was converted into an interactive open access journal. Soon after the transition, the journal experienced a substantial increase of submissions, publications, and citations, demonstrating that traditional journals can be successfully converted into interactive open access journals. Three other open access journals published by EGU (Annales Geophysicae, Natural Hazards, and Earth System Sciences, Non-linear Processes in Geophysics) have maintained traditional peer review up to now. In view of the more successful development of the interactive open access journals, however, they are planning to introduce multi-stage open peer review as well.

The multi-stage open peer review concept of ACP has also been adopted by the e-journal Economics<sup>21</sup> which was launched in 2007 and involves some of the most prominent institutions and scientists in the field of economics. Alternative concepts of public peer review and interactive discussion are pursued by the open access publications Journal of Advances in Earth System Modeling (JAMES; since 2008)<sup>22</sup>, PLoS One<sup>23</sup>, Biology Direct<sup>24</sup>, Electronic Transactions of Artificial Intelligence (ETAI; since 1997)<sup>25</sup>, and Journal of Interactive Media in Education (JIME; since 1996)<sup>26</sup>. Differences between the peer review concepts of these publications and ACP will be addressed and discussed below (see Key features of multi-stage open peer review as practiced by ACP and Comparison to Earlier Initiatives with Two- or Multi-Stage Open Peer Review).

<sup>6</sup> [www.publications.copernicus.org/open\\_access\\_journals/journals\\_by\\_subject.html](http://www.publications.copernicus.org/open_access_journals/journals_by_subject.html)

<sup>7</sup> [www.atmospheric-chemistry-and-physics.net](http://www.atmospheric-chemistry-and-physics.net)

<sup>8</sup> [www.atmospheric-measurement-techniques.net](http://www.atmospheric-measurement-techniques.net)

<sup>9</sup> [www.biogeosciences.net](http://www.biogeosciences.net)

<sup>10</sup> [www.climate-of-the-past.net](http://www.climate-of-the-past.net)

<sup>11</sup> [www.drinking-water-engineering-and-science.net](http://www.drinking-water-engineering-and-science.net)

<sup>12</sup> [www.earth-system-dynamics.net](http://www.earth-system-dynamics.net)

<sup>13</sup> [www.earth-system-science-data.net](http://www.earth-system-science-data.net)

<sup>14</sup> [www.geoscientific-model-development.net](http://www.geoscientific-model-development.net)

<sup>15</sup> [www.hydrology-and-earth-system-sciences.net](http://www.hydrology-and-earth-system-sciences.net)

<sup>16</sup> [www.ocean-science.net](http://www.ocean-science.net)

<sup>17</sup> [www.social-geography.net](http://www.social-geography.net)

<sup>18</sup> [www.solid-earth.net](http://www.solid-earth.net)

<sup>19</sup> [www.the-cryosphere.net](http://www.the-cryosphere.net)

<sup>20</sup> <http://www.egu.eu/publications/open-access-journals.html>

<sup>21</sup> [www.economics-ejournal.org](http://www.economics-ejournal.org)

<sup>22</sup> [www.agu.org](http://www.agu.org), since 2008

<sup>23</sup> [www.plosone.org](http://www.plosone.org), since 2007

<sup>24</sup> [www.biology-direct.com](http://www.biology-direct.com), since 2006

<sup>25</sup> <http://www.etaij.org/>, since 1997

<sup>26</sup> <http://www-jime.open.ac.uk>, since 1996

In short, approaches where interactive commenting and public discussion are not fully integrated with formal peer review by designated referees tend to be less successful.

## FINANCING AND SUSTAINABILITY OF INTERACTIVE OPEN ACCESS PUBLISHING

Atmospheric Chemistry and Physics and its EGU/Copernicus sister journals prove not only the scientific but also the economic viability and sustainability of interactive open access publishing and peer review. The journals were launched and are operated by the independent scientific society EGU and by the small commercial enterprise Copernicus without public subsidies, private donations, or venture capital as involved in the start-up and operation of other successful open access publishers like PLoS and BioMed Central. After several years of operation, ACP and its sister journals have recovered the financial investments of EGU and Copernicus during the start-up phase, and they now deliver a surplus which supports the start-up of new journals by the scientific society as well as a healthy growth of the commercial publisher generating dozens of new jobs.

By developing and applying efficient software tools for the handling of manuscripts (submission, peer review and commenting, typesetting/production, and distribution), and because minimal time and effort is wasted on carelessly prepared papers (high quality of submissions and low rejection rates as detailed above), Copernicus is able to produce top quality publications at comparatively low cost. The publication service charges are of the order of one hundred Euros per page in final double column format, i.e., about one thousand Euros for an average paper with a length of about ten pages. The service charges cover the review support from the editorial office, free use of color figures and online supplementary materials (data, pictures, movies etc.), typesetting of both the discussion and the final version of the paper, archiving and distribution of papers, and interactive comments (maintenance of websites and servers, electronic copies for open archives, paper copies for copyright libraries, etc.) and overheads. In agreement between the publisher (Copernicus) and the scientific society (EGU council and publications committee), the service charges are adjusted to cover the full costs of publishing, including all the tasks and services outlined above, and to generate a modest surplus for the scientific union: ~10% of the annual financial turnover (currently about three million Euros). The surplus is re-invested in publication development (new journals and services) and it helps to run the membership and outreach activities of EGU, which is a non-profit organization. Like the other scientific officers of the union, editors do their work unpaid on a purely voluntary basis. Following up on the questions and suggestions of a reviewer of this manuscript, I would like to clarify that neither I nor any other editor of ACP and the other EGU interactive open access journals have had any income from the journals that we edit as a voluntary community service. In fact, we pay regular registration fees of up to 500 EUR to attend the annual general assembly and scientific conference of our union (EGU), where the editorial board meetings take place. The separation of financial and scientific interests seems important in the context of peer review, and the ACP/EGU experience demonstrates that a purely voluntary approach on the scientific editors' side is sustainable and

compatible with efficient operation of open access journals by a commercial publisher.

For each paper published in ACP, the service charges are levied from the authors or paid by their scientific institution. Since 2008 the German Max Planck Society (MPG)<sup>27</sup> and the French Centre National de Recherche Scientifique (CNRS)<sup>28</sup> have contracts with Copernicus for automated coverage of service charges incurred by their scientists. Other scientific institutions are likely to follow these examples, and many national and international research organizations and funding agencies pursue complementary ways of covering open access service charges for their scientists and projects. Like other open access publishers, Copernicus, and EGU are ready to cover the costs for up to 10% of the papers published each year, if the authors are unable to pay the service charges (e.g., authors without institutional support or institutions from less developed countries). Currently, most papers published in ACP originate from Europe (~50%) and North America (~30%), but the proportion of papers originating from Asia and other regions is increasing.

The ACP open access publication service charges compare quite favorably with the charges levied by other comparable scientific journals and publications:

1. Other major open access publishers such as BioMed Central and the Public Library of Science (PLoS) typically charge more than 1,000 EUR for traditional single-stage journal publications.
2. Traditional publishing groups like Springer charge 2,000 EUR for making individual publications in traditional subscription journals freely available online ("open choice"), i.e., they levy 2,000 EUR per online open access paper in addition to charging libraries and other subscribers for access to the journal in which it appears.
3. In the traditional scientific publishing business, where some journals do not only limit access to subscribers or sell articles on a pay-per-view basis but also request additional publication charges from authors (up to several hundred US dollars per page or color figure), the total turnover, and public costs amount to several thousand US dollars per paper. The annual turnover of journal publishing in the sector of science, technology, and medicine (STM) amounts to around seven billion USD per year, and some of the traditional publishers – led by Elsevier with a market share of about 30% – make operating profits of up to 30% and more. Note that a large proportion of the turnover and profit in STM publishing comes from packaging and selling publicly funded research results that are peer reviewed by publicly funded scientists to publicly funded institutions of education and research (Economist Academic Publishing, 2011; Golden and Schultz, 2012).

In view of these facts, ACP authors and the ACP scientific community have had little difficulty in accepting and paying average service charges of about one thousand Euros per paper to make ACP and its sister journals sustainable. Overall, ACP and its

<sup>27</sup> [www.mpg.de](http://www.mpg.de)

<sup>28</sup> [www.insu.cnrs.fr/](http://www.insu.cnrs.fr/)

interactive open access sister journals prove that top quality (interactive) open access publishing and peer review can be realized and sustained by scientific societies and (small) commercial publishers with tightly limited budgets and without public subsidies, private donations or venture capital. Indeed, ACP, EGU, and Copernicus demonstrate how STM publishing at large can and will hopefully soon manage a swift transition from the past of print-based subscription barriers into the future of an internet-based open access environment.

### KEY FEATURES OF MULTI-STAGE OPEN PEER REVIEW AS PRACTICED BY ACP

The following key features of the ACP multi-stage open peer review system help ensure maximum efficiency of scientific exchange and quality assurance, making it more successful than most other forms of closed or open peer review:

1. Publication of discussion papers before full peer review and revision: free speech, rapid publication, and public accountability of authors for their original manuscript foster innovation and deter careless submissions.
2. Integration of public peer review and interactive discussion prior to final publication: attract more comments than post-peer review commenting, enhance efficiency, and transparency of quality assurance, maximize information density of final papers.
3. Optional anonymity for designated referees: enables critical comments and questions by referees who might be reluctant to risk appearing ignorant or disrespectful – especially when providing a voluntary community service in which they have little to gain for investing lots of effort and time.
4. Archiving, public accessibility, and citability of every discussion paper and interactive comment: ensure documentation of controversial scientific innovations or flaws, public recognition of commentators' contributions, and deterrence of careless submissions.

Combining all of the above features and effects is the basis for the great success of ACP and its sister journals. Missing out on one or more of these features is the main reason why most if not all alternative forms of peer review practised in other initiatives for improving scientific communication and quality assurance have been less successful (less commenting, lower impact/visibility, higher rejection rates, larger waste of refereeing capacities, etc.).

For example, the release of a “pre-publication history” and/or the opportunity for “peer commentary” after completion of the actual peer review and publication of the final revised manuscript as practiced by the BMC medical journals of BioMed Central<sup>29</sup> as well as the journals Behavioral and Brain Sciences<sup>30</sup> and Psychology<sup>31</sup> are very useful advances and improvements compared to traditional journal publishing, but they miss some of the above features and advantages. Controversial scientific innovations or flaws in papers rejected after peer review are not documented for

the public and scientific community. Moreover, the completion of peer review and revision before publication and public discussion of a manuscript does not allow interested members of the scientific community to have any input to the revision and the final editorial decision. Obviously, “post-commenting” after peer review is much less attractive to scientists than commenting in the course of peer review. The latter allows individual scientists to support and influence the conclusions and publications of their colleagues, e.g., by pointing out related earlier findings and studies which the authors can still include in the reference list of the manuscript thus in standard citation analyses. In contrast, post-commenting after final publication does neither enable the commentator to influence the final publication, nor does it allow the authors to improve their publication along the lines suggested by the commentators. Accordingly, potential commentators have not only less incentive to invest effort and time in contributing to their colleagues' and competitors' work; they also have to worry that critical comments might just be regarded as a devaluing critique rather than a helpful contribution. This fairly straightforward consideration is supported by the fact that most journals with post-commenting receive fewer comments from the scientific community (Müller, 2008). For example, only one of ~20 papers published in PLoS One receives a comment from the scientific community (as opposed to one of ~5 in ACP), although PLoS offers more advanced and easier to use commenting tools and tries to advertise and promote the commenting more actively than ACP.

For several reasons also the “open peer review trial” of the Nature magazine in 2006 was not a good example and measure for the engagement of scientists in interactive commenting and public peer review on the internet. In that experiment, neither the authors of an article nor their colleagues and readers had much of an incentive to participate in the public discussion. The authors had to accept that their article was exposed in parallel to public scrutiny as well as to a closed peer review process where the referee comments remain non-public and where most of submitted manuscripts are rejected not because of a lack of scientific quality but because they are not deemed sufficiently exciting for the interdisciplinary audience of the magazine (ca. 93% rejection rate)<sup>32</sup>. For the likely outcome that a manuscript would not pass the closed peer review, it was not clear whether and in which form the rejected manuscript and the public comments would remain publicly accessible. As one might have imagined beforehand, this is not a very attractive perspective for scientists trying to get recognition for their most exciting results. Similarly, colleagues and readers had little incentive to formulate and post substantial comments, because their contributions would just have been an addendum to the closed peer review proceeding in parallel and would likely disappear afterward. Fortunately, the publishers of Nature seem to have realized that permanent archiving and citability are key features of scientific exchange, and they have launched a more promising initiative titled Nature Precedings. There manuscripts can be published, openly discussed and archived in a similar way as in the discussion forums of interactive open access journals<sup>33</sup>.

<sup>29</sup>www.biomedcentral.com

<sup>30</sup>www.bbsonline.org

<sup>31</sup>psycprints.ecs.soton.ac.uk

<sup>32</sup>www.nature.com/nature/peerreview/debate/nature05535.html

<sup>33</sup>http://precedings.nature.com/site/help

Unfortunately, however, it seems that the paramount importance of archiving and citability of manuscripts and comments has not yet been fully recognized by scientific publishers and societies. Following up on the success and leadership of the EGU in interactive open access publishing and peer review, the American Geophysical Union (AGU) has recently also engaged in experiments with “open peer review.” Instead of building on the very positive experience and success of the European sister society, however, AGU seems to follow the tracks of the unsuccessful earlier trial of Nature. Specifically, AGU announced that the discussion paper and all interactive comments shall be deleted after completion of the peer review process and final acceptance or rejection of the revised manuscript (Albarede, 2009). This line was also followed in the JAMES, which had originally adopted the interactive open access journal concept of ACP but then abandoned the archiving of discussion papers in their discussion forum (JAMES-D) and was recently taken over by AGU. If AGU were to continue the approach of erasing discussion papers and comments, they would largely miss out on the effects detailed under point 4 above, and it appears questionable that the perspective of deletion after a couple of months will attract substantial commenting from the scientific community. Hopefully, the proponents of the AGU experiment will realize that the deletion of scientific comments is not only a discouragement for potential commentators but also a regrettable underestimation of the value of scientific discussion and discourse in the history and progress of science.

As outlined on the web pages of ACP/EGU, the permanent archiving of discussion papers can occasionally lead to inconveniences for authors and other parties involved in the review and publication process. Overall, however, the advantages of permanent archiving clearly outweigh the potential disadvantages<sup>34,35</sup>.

For the following reasons it would be neither appropriate nor possible to delete discussion papers after they have been published online:

- (1) The deletion of published materials is incompatible with the virtues of traceability and reliability that are central to science and scientific publishing in general, and to the interactive open access publishing approach of ACP/EGU in particular. Deleting published scientific information is against the very nature of science.
- (2) The deletion of discussion papers and comments would discourage potential commentators, and it would imply a disregard for the value of scientific discourse (Pöschl, 2010b, pp. 305–306).
- (3) The use of digital object identifiers (DOI) entails legal obligations of ensure permanent archiving and accessibility.
- (4) Even if it were desirable, it would be practically impossible to “unpublish” a discussion paper published in ACPD. Upon online publication, the papers are copied into multiple electronic repositories. Moreover, referees, readers, and

other internet users can and do download copies for storage at arbitrary locations that are beyond the control of any the publisher. Therefore, a published paper can be formally withdrawn/retracted by publication of a commentary analogous to stating the reasons like in traditional print journals. It can, however, not be “unpublished” by deletion from the web pages and archives of the journal.

One of the central aims of interactive open access publishing is high efficiency in scientific communication and quality assurance. As detailed in the attached articles, the average quality and visibility of ACP, and its sister journals are higher than those of most comparable journals while the rejection rates are lower. The highly efficient mechanism of review, publication, and self-regulation would hardly work if authors could submit manuscripts at any rate and simply delete published discussion papers if the public peer review and editorial decision were not favorable (or for any other reason).

Experience and rational thinking suggest that multi-stage open peer review should be applicable and beneficial for journal publications in most if not all disciplines of scientific research (STM as well as social sciences, economics, and humanities). For consistency and traceability, discussion papers, and interactive comments should generally remain archived and citable as published, and they should be regarded as proceedings-type publications. Due to the proceedings character of discussion papers, the authors of revised manuscripts that may not have been accepted for final publication in the interactive open access journal to which they had originally been submitted can still pursue review and publication in alternative journals. As indicated above, such aspects are particularly important with regard to highlight magazines or journals in which the review process is not only aimed at ensuring scientific quality but also at high selectivity with regard to interdisciplinary relevance and visibility, which entails low probability of acceptance even for manuscripts of high quality (see Nature trial).

In addition to the above general features, the following specific procedural aspects have turned out to be important for the practical implementation and effectiveness of interactive open access publishing and peer review:

#### EDITOR ASSIGNMENT

For the assignment of a newly submitted manuscript to a handling editor, the online editorial office automatically sends invitation letters to all editorial board members covering the relevant subject areas (based on index terms selected by authors). Depending on competence and availability, each editorial board member can then decide if s/he wants to take editorship (first come, first served; every board member is expected to handle at least six submissions per year). If no handling editor can be found via the automated assignment process, the authors and the executive editors are informed and asked to directly contact individual board members if they are ready to take editorship. This second line of editor assignment in ACP is similar to the regular editor assignment procedure in the open access journal Biology Direct<sup>36</sup>. There it is up to the

<sup>34</sup>[http://www.atmospheric-chemistry-and-physics.net/general\\_information/faq.html](http://www.atmospheric-chemistry-and-physics.net/general_information/faq.html)

<sup>35</sup><http://www.egu.eu/statements/position-statement-on-the-status-of-discussion-papers-published-in-egu-interactive-open-access-journals-4-july-2010.html>

<sup>36</sup>[www.biology-direct.com/info/about/](http://www.biology-direct.com/info/about/)



authors to find and motivate an editorial board member to guide the review process for their paper, and the manuscript is effectively rejected if none of the board member agrees.

### ACCESS REVIEW

Prior to publication in the discussion forum, the editor is asked to evaluate whether the submitted manuscript is within the scope of the journal and whether it meets basic quality criteria. If necessary, the editor may consult referees for a rapid and preliminary initial rating of the manuscript<sup>37</sup>. The editor or referees can request/suggest minor technical corrections and adjustment (typing errors, clarifications, etc.). Further requests for revision of the scientific contents are not allowed at this stage of the review process but shall be expressed in the interactive discussion following publication of the discussion paper. For rapid processing and in order to save refereeing capacities the editor shall normally perform the access review without the referees, unless their advice is urgently needed or the authors have requested their involvement. In a statement or cover letter accompanying the submitted manuscript, the authors can indicate if they have any preference on involving the referees already in the access review. Obviously, the involvement of referees can lead to delays, but on the other hand the authors may want to receive a preliminary rating and suggestions for minor corrections prior to publication of the discussion paper.

### FINAL RESPONSE AND REVIEW COMPLETION

In the final response phase at the end of the interactive public discussion, the authors shall respond to all comments. The editor has the opportunity of adding comments and suggestions, but normally editorial decisions and recommendations should not be taken and expressed before the authors have responded to all comments (“*audiat et altera pars*”). Instead, it shall be up to the authors to decide if they want to pursue final publication and how they shall revise their manuscript in view of the public review and discussion (self-regulation once again). Depending on the situation, they can but need not ask and wait for the editor to give advice on how to proceed and if a revised version is likely to be accepted for final publication. After receiving critical feedback, mature, and responsible scientists should normally know best how to revise their manuscript. Indeed, the improvements upon revision of a manuscript after public discussion often go far beyond the requests and suggestions expressed by the referees. Premature interference by the editor would likely reduce rather than enhance the authors’ motivation for improving the manuscript upon revision. Moreover, premature editorial recommendations published by the editor before seeing the authors’ final response and the revised manuscript could potentially bias the final decision about acceptance or rejection.

After receiving the revised manuscript the editor has a complete picture, can check if all comments and suggestions have been properly taken into account, and can suggest or request further improvements. If required, the process of review and revision can be iterated with the help of referees. So far, such iterations of

peer review as well as appeal procedures in case of controversial editorial decisions have not been handled in public to avoid unnecessary complications. In the end, however, the discussion forum can and shall be used to explain editorial decisions in a rational and transparent way as illustrated by the following example (Pöschl, 2009b)<sup>38</sup>:

Currently, the editorial guidelines of ACP encourage editors to publish scientifically useful referee-author exchanges from non-public part of peer review completion in similar ways as in the exemplary case cited above. In the future, intermediate manuscript versions and related comments from the access review or the review completion shall be automatically made available upon publication of a manuscript in ACPD or ACP, respectively (analogous to pre-publication history available in BMC medical journals). If, however, a newly submitted manuscript is not accepted for publication in ACPD or a revised manuscript is not accepted for publication in ACP, the manuscript, and related comments shall be kept confidential in order to avoid escalation of scientific disputes and to maintain the authors’ opportunity of pursuing publication in alternative publishing venues (European Geosciences Union, 2010).

### COMPARISON TO EARLIER INITIATIVES WITH TWO- OR MULTI-STAGE OPEN PEER REVIEW

Following up on the requests of a referee in the peer review of this manuscript, the following paragraphs provide a detailed comparison to earlier initiatives with similar concepts and a discussion of potential reasons for different developments. During the initiation and planning of ACP and its interactive journal concept in the years 2000 and 2001, I was looking for – but was unable to find – similar initiatives to compare with and learn from (Pöschl, 2004). It was only at an e-publishing workshop of the Max Planck Center for Information Management in May 2002 that I learned of a similar initiative launched as early as 1996: the JIME<sup>39</sup>. Coming from a completely different scientific background, the founders of JIME had designed and realized a similar concept of multi-stage open peer review with public discussion. Unfortunately, however, JIME attracted only a small number of publications and seems not to have inspired the foundation of similar journals in related fields of science and humanities. Despite the overall conceptual similarities, JIME does not show some of the key features of the ACP interactive journal concept. In particular, the “private open peer review” of JIME foresees a non-public exchange of arguments between referees and authors, which is opened to the public only after approval by the editor. This seems to limit the publication and documentation of controversial scientific innovations or flaws much more than the “access peer review” of ACP (quick go/no-go decision essentially without non-public exchange of arguments between authors and referees). Moreover, all referees are named and no anonymous referee comments are allowed in JIME, which is likely to limit and inhibit critical review and discussion. These differences may appear subtle at first sight, but they are highly relevant for the practical operation of a scientific journal and may be decisive for its success and acceptance in the target scientific community.

<sup>37</sup> [www.atmospheric-chemistry-and-physics.net/review/ms\\_evaluation\\_criteria.html](http://www.atmospheric-chemistry-and-physics.net/review/ms_evaluation_criteria.html)

<sup>38</sup> [www.atmos-chem-phys-discuss.net/8/S12406/2009](http://www.atmos-chem-phys-discuss.net/8/S12406/2009)

<sup>39</sup> [www.biomedcentral.com](http://www.biomedcentral.com)



After JIME, I got to know about another early online publication format with a two-stage open peer review process: the ETAI<sup>40</sup> launched in 1997. Similar to JIME, ETAI attracted a series of special issues related to conferences or projects, but the number of individually submitted articles remained small. Regular operations stopped in 2002, but the ETAI home page indicates plans for a re-launch. As described by Sandewall (1997, 2006, 2012), the open peer review process of ETAI does not integrate but separate the two major aims of peer review, namely, to improve the quality of submitted manuscripts and to establish certain quality standards. The first stage is an interactive public discussion which invites questions, comments, and suggestions from the scientific community, but it does not involve designated referees, and all participants are openly named. In a second stage, anonymous referees decide about acceptance of the revised manuscript for ETAI, and further rounds of revision are normally not allowed. These features of ETAI bear similarities to the unsuccessful trial of open peer review by Nature magazine in 2006, and they are in stark contrast to the ACP review process, where the referees contribute to the interactive public discussion and have an option of staying anonymous, and the peer review process can be continued iteratively like in traditional journals. For the authors and readers of ETAI it seems not clear, if the openly named participants of the interactive public discussion in the first stage of the review process might also serve as an anonymous referees in the second stage. It seems rather unattractive for authors to post their manuscript for open discussion and scrutiny by the scientific community, and to have only one chance of revision before anonymous referees who may or may not have been involved in the preceding discussion are expected to make a “pass/fail decision” (Sandewall, 2012). In the relatively few review processes that have actually been completed in ETAI so far (several dozens in the time frame of 1997–2002), all involved parties seem to have requested exceptions, i.e., anonymity in the interactive public discussion and iterative revisions in the second stage of review (Sandewall, 2012). Both of these “exceptional” features are key elements of the successful ACP approach. From long-term experience with several thousand review processes completed in ACP since 2001, we know that these features are vital for the large success of the EGU interactive open access journals, and I would argue that they might be critical for the limited success of ETAI. In any case, the ACP/EGU approach of multi-stage open peer review is aimed at integrating rather than separating the processes of interactive public discussion and classical peer review as well as the aims of manuscript improvement and quality control.

The limited success of JIME and ETAI compared to ACP demonstrates the difficulties of practical implementation and the importance of the conceptual aspects and subtleties outlined above (see Key features of multi-stage open peer review as practiced by ACP). Nevertheless, the basic aims and principles of JIME, ETAI, and ACP are similar, and their independent development in different disciplines including the social, natural, and computer sciences reflects the power of the idea and the appeal of transparency in scientific quality assurance.

The review article of Sandewall (2012) outlines and compares further analogies and differences between ETAI and ACP, and it also provides a very useful and comprehensive account of challenges faced by proponents of open peer review. In the following paragraphs I am following up on some of the questions and issues raised.

- (1) Defining different types of scientific publication (Sandewall, 2012: p. 2–3): Robust and self-consistent definitions of different types of scientific publications are indeed important for scientific communication and quality assurance. I would, however, not tie such definitions to electronic vs. non-electronic or different types of publishers. Instead, I would suggest to use self-explanatory terms that are meaningful regardless of the publishing medium. Along these lines, the term “discussion paper” has proven well defined and useful as specified in a position statement of the EGU with references to other scientific societies and publishers. Thus, I would recommend broad usage of this term for the first stage of publication in two- or multi-stage open peer review.
- (2) Resolving doubts about the viability of open peer review (Sandewall, 2012: Section 4.1): For the reasons outlined by Sandewall (2012) it is important to demonstrate the viability and advantages of open peer review with practical examples. The statistics of ACP and its sister journals prove that the arguments given in Section 4.1 of Sandewall (2012) are valid and applicable to a wide range of research areas involving scientists trained in physics, chemistry, biology, geology, engineering, and other disciplines. Besides a clear concept and terminology (“discussion paper,” etc.), it is important to have a dedicated team of scientists who do not only advertise and explain the new approach but also demonstrate its practical viability by submitting and publishing high quality papers (see below).
- (3) Starting the flow of submissions and debate (Sandewall, 2012: Section 4.2): Starting a steady flow of submissions is indeed the most important task for the editorial board of any new journal – even more so for an innovative journal experimenting with new forms of peer review. In most areas of natural science, a journal can be regarded as well established only when it is covered by major indexing services and acquires a journal impact factor or equivalent measure of visibility, which usually takes at least a couple of years. Until then, colleagues without genuine interest in the journal cannot be expected to submit high quality manuscripts that would likely reach higher visibility and citation counts elsewhere. Thus, it is up to the editorial board members and other supporters to maintain a steady flow of high quality submissions. For this purpose as well as for efficient handling of manuscripts when the flow of submissions increases, it is helpful to gather a large editorial board that is firmly rooted in the scientific community and includes experts for all subject areas of the journal scope (ACP: ~70 board members at the beginning, ~130 now).

Initiating the review and discussion of manuscripts with high quality comments that set a precedent for further commenting is of course also important for journals with open peer review. In ACP and its interactive open access sister

<sup>40</sup>[www.insu.cnrs.fr/](http://www.insu.cnrs.fr/)

journals this is mostly done by designated referees appointed by the editor handling the submission. Unsolicited comments can be expected only if members of the scientific community have a strong interest to ask for more information or suggest corrections/additions concerning the methods, results, and conclusions of a study. As expected, non-controversial papers usually receive comments only from the designated referees. Other scientists have little incentive to invest effort and time commenting on papers that they may find potentially useful but not controversial.

- (4) Maintaining coherence (Sandewall, 2012: Section 4.3): For ACP, coherence is not more of an issue than for traditional journals covering multiple subject areas with the help of multiple editors. The journal scope has to be well defined and reflect the interests and quality standards of the scientific community served by the journal. Different communities tend to have different standards and preferences with regard to both the format and the content of manuscripts. Therefore, EGU publishes multiple topical journals rather than just one large geosciences journal including all disciplines. Even within the discipline of atmospheric science, EGU publishes more than just one journal, namely ACP and the sister journal “Atmospheric Measurement Techniques” (AMT) which is focused on method development and exhibits similarly high growth rates of volume and visibility as ACP. Due to the transparency of the review process and related self-regulation mechanisms, the quality of final papers published in ACP is generally not more variable than in traditional journals with smaller editorial boards. The ACP editors do not spend extra time on moderating the interactive public discussions, which are not actively moderated for the reasons outlined above. Compared to traditional journals where the editors often rely on simple majority votes of the referees, however, the ACP editors tend to spend more time on carefully validating the referee recommendations, because the transparent review process publicly reveals editorial decisions that are not well-founded.
- (5) Computational and administrative infrastructure (Sandewall, 2012: Section 4.4): The installation and maintenance of computational and administrative infrastructure is the main reason why the operation of an open access journal is not cost free, even if most of the review work is done by volunteers. The referees and editors of EGU journals receive no financial rewards. The editors even pay the regular registration fee to participate in the annual EGU General Assemblies with over 10,000 participants where the editorial board meetings take place. The small commercial publisher Copernicus is a spin-off from the Max Planck Society and continues to aim for providing optimal infrastructure and services at minimal cost. Nevertheless, it seems difficult to reduce the average costs far below one thousand Euros per paper, but this is anyhow much lower than the prices of most traditional publishers as discussed above (see Financing and Sustainability of Interactive Open Access Publishing).
- (6) Maintaining liveliness of peer review discussion (Sandewall, 2012: see Comparison to Earlier Initiatives with Two- or Multi-Stage Open Peer Review): For the reasons outlined by Sandewall (2012), it is difficult if not impossible to ensure a lively review discussion for all papers published in large scientific journals. This may be problematic for the two-stage review approach of ETAI, where the first stage is designed as a pure community discussion without the involvement of designated referees. For the integrative approach of ACP, however, it is not problematic that most papers receive comments only from the designated referees. The transparency of the peer review process and the option for additional input from the scientific community are sufficient to stimulate self-regulation and enhance the efficiency of scientific quality assurance (Pöschl, 2004, 2010a,b). Discussion papers that report controversial findings often do attract unsolicited comments from the scientific community, but why would researchers invest effort and time in the commenting of their colleagues’ publications which they may find interesting but not controversial? Sometimes more commenting and discussion might be useful, but usually the volume of comments exchanged between authors and referees amounts to as much as 50% of the discussion paper volume, and further commenting can be cumbersome – especially for the authors who normally do not want to spend too much time and effort on the discussion of a single paper but rather move on to the next study. Therefore, unnecessary comments and artificial liveliness of discussion might actually deter authors and do more harm than good to a journal with open peer review.
- (7) Open names policy (Sandewall, 2012: Section 7.1): In an ideal world, where people generally react positive to criticism and where scientists can dedicate unlimited amounts of time and effort into compiling completely accurate reviews about their colleagues’ manuscripts, I would agree that referee anonymity should be abandoned. In practice, however, optional anonymity for referees appears appropriate or even necessary to enable critical comments and questions by referees who might be reluctant to risk appearing ignorant or disrespectful (Pöschl, 2004). As outlined above, less than 20% of the referee comments published in the discussion forum of ACP are posted with the name of the referee, i.e., the referees prefer in most cases (>80%) not to reveal their name. Purists often suggest that offering anonymity to referees would be unfair against the authors of a manuscript, and that both parties should be openly named to ensure equal rights and opportunities. They tend to forget, however, that the authors want to get their paper approved by peers, and that the referees usually provide this service on a voluntary basis. In this sense, the authors actually exploit the working capacities of the referees, and the peer review process offers a major gain to the authors (conversion of their manuscript into a peer reviewed paper) but relatively little benefit to the referees. Therefore, it seems appropriate to protect the referees from potential negative consequences of the free service they provided to the authors and to the scientific community. The very small number of author complaints about inappropriate referee comments (about one in 10,000) and the low rejection rates of manuscripts submitted for peer review in ACP and the other EGU interactive open access sister journals (generally <15%) confirm that transparency of the review process (open-process peer review) is normally sufficient to

protect authors from inappropriate referee comments. Thus, it seems neither necessary nor appropriate to abandon optional anonymity, impose an open names policy and force referees to reveal their identity. All available evidence suggests that refereeing capacities are the most limited resource in scientific publishing and quality assurance (Pöschl, 2004). In view of the ever-increasing flow of manuscripts submitted for peer reviewed publication, it appears more important to protect referees rather than authors – especially in a multi-stage open peer review process like that of ACP, where the authors anyhow have the opportunity of free speech through their discussion paper and the interactive comments they can post during the open discussion as well as in a final response phase where no more referee comments are allowed<sup>41</sup>.

### KEY QUESTIONS FOR OPEN EVALUATION IN SCIENCE

The coordinators of the special issue hosting this article posed a series of ten key questions to be considered in designing and implementing a concept of open evaluation in science. More than a decade of practical experience and success in re-shaping the processes of scientific publishing and quality assurance as well as continued exchange with scientists and publishing professionals from various disciplines in the sciences and humanities lead to the following answers.

- (1) Should some evaluation take place prior to publication or should all evaluation occur post-publication? Experience and rational consideration suggest that the main review process should take place before (final) publication of a manuscript. A fundamental disadvantage of pure post-publication review is that the reviewers cannot contribute to a revision and improvement of the published manuscript. Thus, both the authors and the reviewers are likely to consider critical comments as destructive rather than constructive. Moreover, the reviewer has less incentive to invest effort and time in suggesting additions and corrections, including but not limited to referencing relevant related publications. Last but not least, post-publication commenting does not enhance the information density of scientific communication. If the reviewer comments cannot be implemented in a revised manuscript, the readers have to consult all comments and extract the information from there, which is much less efficient than reading a revised manuscript that synthesizes the information exchanged in the review process. For the above reasons, most publishing platforms that offer only post-publication commenting attract rather small numbers and volumes of comments.
- (2) Should reviews and ratings be entirely transparent, or should some aspects be kept secret? Reviews and ratings pertaining to a published manuscript should be made entirely transparent. Reviews and ratings of manuscripts that do not achieve (final) publication, however, should be kept confidential to avoid public escalation of scientific disputes and to give authors a chance of pursuing publication of their (revised) manuscript in alternative publishing venues.
- (3) Should alternative metrics, such as paper downloads be included in the evaluation? Paper download statistics are among the many possible forms of post-publication evaluation and should certainly be considered for comprehensive evaluation of scientific publications, but not without precautions against manipulation and misinterpretation of this relatively primitive usage metric. Many scientific journals, including traditional subscription journals with hidden peer review, are already providing download data and highlighting most downloaded papers. This approach certainly facilitates the detection of “hot papers,” but compared to long-term citation statistics and other usage metrics it seems less robust and should not be overrated.
- (4) How can scientific objectivity be strengthened and political motivations weakened in the future system? Like in all branches of human society and politics, transparency, and free speech appear to be the best if not the only sustainable way of pursuing objectivity in a balance of powers and interests.
- (5) Should the system use signed and authenticated reviews and ratings or anonymous ones, or both? An entirely open and traceable exchange of scientific arguments in the form of signed and authenticated comments is certainly desirable and shall be encouraged. For practical reasons, however, it seems appropriate and beneficial to allow also for anonymous reviews. Optional anonymity enables critical comments and questions by referees who might be reluctant to risk appearing ignorant or disrespectful – especially when providing a voluntary community service in which they have little to gain for investing lots of effort and time.
- (6) Should the evaluation be an ongoing process, such that promising papers are more deeply evaluated? The evaluation of scientific publications has to be and generally is an ongoing process – with regard to citation counting as well as commenting and other forms of evaluation that are and have long been in use. Note that also traditional journals with hidden peer review also allow for commentaries referring to earlier papers. In practice, however, relatively few papers seem to attract comments after (final) publication. Moreover, most authors seem to prefer finalizing a publication at some point, and following up with new studies rather than continuously revising and updating old papers. For certain types of publications such as review articles, continuous extension, and revision may be a good and attractive approach as exemplified by the Living Reviews project and journal family<sup>42</sup>. For standard articles presenting new scientific findings, however, a finite process of publication appears more straightforward. Either way, thorough evaluation of scientific studies seems difficult if not impossible without long-term perspective.
- (7) How can we bring science and statistics to the evaluation process (e.g., should rating averages come with error bars)? Scientific reviews and ratings are necessarily subject to the

<sup>41</sup>[http://www.atmospheric-chemistry-and-physics.net/review/review\\_process\\_and\\_interactive\\_public\\_discussion.html](http://www.atmospheric-chemistry-and-physics.net/review/review_process_and_interactive_public_discussion.html)

<sup>42</sup><http://www.livingreviews.org/>

same uncertainties and progress as the studies that undergo rating and review. Thus, it seems natural to assess also the reliability of reviews and ratings. One of the many advantages of open peer review is the public availability of reviews and ratings, which makes them accessible for statistical analysis. Thus open access and open peer review inherently promote the development of new and improved evaluation metrics – in analogy to traditional indexing services like the ISI Web of Science and Elsevier's SCOPUS, but much more efficiently and comprehensively because of unrestricted access and free competition for optimal solutions.

- (8) How should the evaluative information about each paper (e.g., peer ratings) be combined to prioritize the literature? The combination and balancing of different types of evaluative information (ratings/reviews, download/citation statistics, and other usage metrics) will necessarily depend on the aims and perspectives of different types of evaluation or prioritization. For example, the criteria of an evaluation exercise will likely differ for individuals and institutions, scientific researchers and teachers, innovation, and reliability, short-term and long-term impact, etc. In any case, it should be kept in mind statistical indicators are sometimes useful but always also prone to misinterpretation (see publication and citation counting, impact factors, h-indices, etc.).
- (9) Should different individuals and organizations be able to define their own evaluation formulae (e.g., weighting ratings according to different criteria)? Obviously, different individuals and institutions may pursue different goals and should thus be able to apply different criteria and weighting schemes. Moreover, evaluators and service providers should compete in developing the best possible metrics and indicators. This is already the case with ISI Web of Science and Elsevier SCOPUS, and through open access and open peer review many more parties can participate, contribute, and help to overcome the obsolete monopoly/oligopoly structures of scientific indexing.
- (10) How can we efficiently transition toward the future system? An efficient transition to open evaluation in science can be achieved by combining the strengths of traditional peer review with the opportunities of interactive and transparent community assessment on the internet. The concept of multi-stage open peer review has been designed and successfully applied to induce this transition in the geosciences and is spreading into other disciplines. It can be flexibly adjusted to the needs and peculiarities of different scientific communities, and it has the potential of replacing hidden peer review as the standard of scientific quality assurance and forming the basis of an open evaluation system.

## CONCLUSIONS AND OUTLOOK

ACP and its sister journals very clearly demonstrate that interactive open access publishing with a multi-stage peer review process effectively resolves the dilemma between rapid scientific exchange and thorough quality assurance. They have proven that multi-stage open peer review indeed fosters scientific discussion, deters submission of sub-standard manuscripts, saves refereeing capacities, and enhances information density in final papers. Moreover, ACP,

EGU and Copernicus prove the financial sustainability of open access publishing, and they may serve as a role model for how STM publishing at large can manage the transition from the past of print-based subscription barriers into the future of internet-based open access. The key for a successful, smooth, and efficient transition is to utilize the opportunities of modern technology and interactivity while maintaining the strengths of traditional structures and procedures.

Multi-stage open peer review easily can be integrated into new and existing scientific journals as well as large-scale publishing systems and repositories such as arXiv.org – simply by adding an interactive discussion forum. Equipped with appropriate interactive commenting tools, a large repository such as arXiv.org could not only serve as an archive for “preprints” or “e-prints,” but also as a platform for efficient review and discussion, where authors could post their discussion papers and different journals could send their referees for public review. Similarly, individual publishers could set up central discussion forums to serve different journals or journal sections (Pöschl, 2004, 2010b). This perspective is in line with the selected papers network concept of Lee (2012) and the decoupled journal concept of Priem and Hemminger (2012). Depending on the outcome of public review and discussion, the revised manuscripts could then be sorted and grouped at different levels of relevance for different audiences – analogous to the quality ranking system and tiers of the Berkeley Electronic Press journals in economics<sup>43,44</sup>. Another feature that could be integrated in multi-stage open peer review is a double-blind approach in the initial access review (pre-screening) to avoid/minimize bias in selection of discussion papers. In the open discussion, however, it seems more useful and efficient to discuss openly without hiding identities (except for protecting referees if they wish to stay anonymous).

For interdisciplinary highlight papers, EGU and Copernicus are currently preparing the introduction of a third stage of interactive open access publishing that shall lead to efficient grouping of scientific publications in three tiers with the following characteristics:

1. Discussion forum (discussion papers and interactive comments):
  - free speech (for authors and scientific community)
  - original opinions
  - immediate publication and dissemination
2. Topical journal (final papers):
  - thorough quality assurance (collaborative peer review)
  - comprehensive, complete and validated information
3. Highlight magazine (abstracts):
  - highly condensed information
  - interdisciplinary relevance and public interest
  - three-stage selection process (distillation).

The interactive open access highlight magazine shall be dedicated to the selection and presentation of the abstracts of highlight papers, which outline the forefront of research and are of high

<sup>43</sup>www.bepress.com/bejm

<sup>44</sup>www.bepress.com/bejte

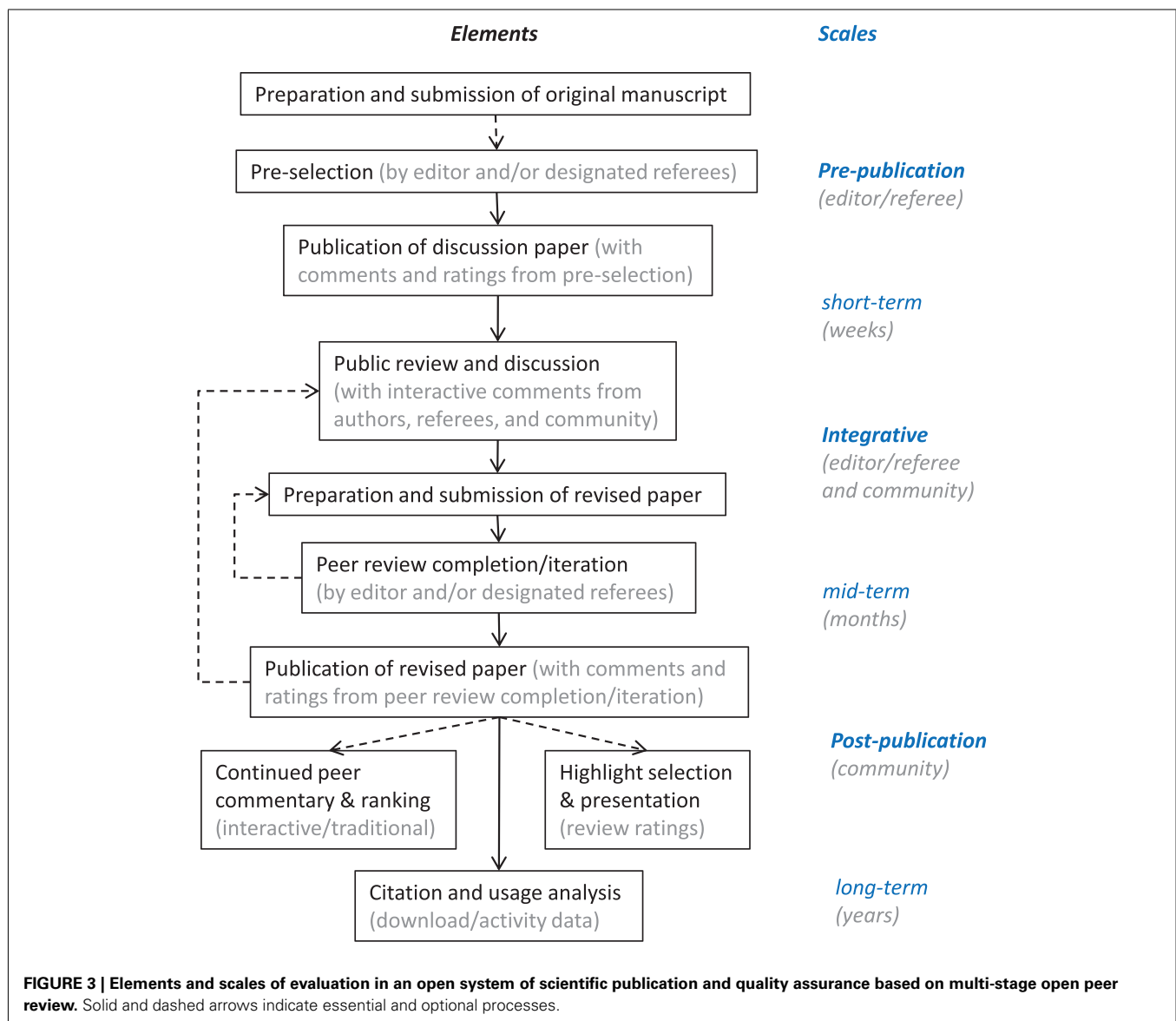
interdisciplinary relevance and public interest. The editorial board of the magazine shall select highlight papers that have undergone public peer review and discussion in topical open access journals, and the abstracts of the highlight papers shall be commented and compiled with direct references and links to the original papers and journals, respectively. By building on rather than competing with topical scientific journals, the highlight selection process and magazine shall provide high efficiency, conciseness and inter-disciplinarity without compromising scientific completeness and quality assurance. This might also be a way forward for traditional highlight magazines like Nature or Science covering the full width of scientific disciplines.

The basic concepts of interactive open access publishing and peer review can be easily adjusted to the different needs and capacities of different scientific communities by maintaining or abandoning referee anonymity, shortening, or prolonging the public discussion phase, adding post-peer review commenting and

rating tools for readers, making all steps/iterations of peer-review and revision transparent, adding further stages of publication for re-revised manuscripts, establishing feedback loops for editorial quality assurance, etc.

**Figure 3** illustrates essential elements and scales of evaluation in an open system of scientific publication and quality assurance based on multi-stage open peer review. While much of the general discussion about reforming scientific quality assurance and evaluation is focused on a distinction of pre- and post-publication processes, the experience and achievements of ACP and EGU show that an integrative approach combining pre- and post-publication elements in a multi-step process of review and publication is most efficient.

Besides communication and evaluation of scientific results, multi-stage open peer review might also be applicable for efficient evaluation of scientific research proposals in the form of citable discussion papers. Again all involved parties could profit





from public documentation, scrutiny, and citability. At first sight, it might appear that the authors of a proposal would run a high risk of “losing” innovative project ideas to the public. In practice, however, they might be better protected from (hidden) plagiarism and obstruction by competitors, and the citable publication might actually help them to claim authorship, precedence, and recognition for their ideas. At the same time, the scientific community and society at large might profit from rapid dissemination of innovative ideas.

Overall, interactive open access publishing and peer review can strongly enhance scientific exchange and quality assurance. The concept has been very successfully applied and extended over the past decade, demonstrating both the scientific benefits and the financial sustainability of open access. It will likely emerge as a best practice model for the future of scientific publishing, and it provides a solid basis for efficient use and augmentation of scientific knowledge in the global information commons (David and Uhler, 2005). Moreover, public review, discussion, and documentation of the scientific discourse can serve as an example for rational and transparent procedures of settling complex questions, problems, and disputes. It is a model for further development of the structures, mechanisms, and processes of communication and decision making in society and politics in line with the principles of critical rationalism and open societies.

A major limiting factor for the development of innovative scientific publication and evaluation systems is the scarcity of funds specifically dedicated to covering open access publication costs. Nevertheless, more and more funding agencies do provide funds for this purpose, and the success of the EGU/Copernicus as well as other open access publishers shows that many scientists are willing and able to cover the costs of open access publishing via publication fees. Overall, the money required to produce scientific publications in a format that is accessible via the internet is already in circulation. Otherwise, the publishers would not be able to offer online subscriptions. Currently, however, the funds are channeled through a rigid subscription system, which has the consequence that certain publishers can make excessively large profits and that the scientific information remains locked away. If the same amount of money were channeled through a flexible open access funding scheme, the same products (scientific journals and papers) could be produced and made freely available on the internet at the same or lower cost in a proper publishing market rather than the current subscription scheme with oligopoly character.

In order to accelerate the improvement of scientific communication and evaluation in a global information commons, I would like to renew the following propositions and recommendations to scientists and scientific publishers, librarians, institutions, and funding agencies (Pöschl, 2004, 2010b):

1. Promote open access to publicly funded research publications by appropriate guidelines and by moving funds from subscription budgets to publication budgets – preferably at high rates (20% per year or more). Obviously, traditional publishers are reluctant to undermine their profits as long as they can rely on rigid subscription schemes, but the ones who are ready to serve science will swiftly adapt to new financing schemes as illustrated by the open choice model and acquisition of BioMedCentral by Springer<sup>45</sup>. The others can be substituted by new service providers as indicated by the swiftly growing number, size, and visibility of open access publishers and journals<sup>46,47</sup>.
2. Promote multi-stage open peer review in new and existing journals, repositories, and other publication platforms. Public review and interactive discussion are technically straightforward and can be flexibly adjusted to different scientific communities, but care should be taken when dealing with key features of peer review and scientific discourse (optional anonymity for designated referees, permanent archiving, and citability of published manuscripts and comments, etc.).
3. Promote the development and implementation of new and improved metrics for the impact and quality of scientific publications (combination of citation, download/usage, commenting, and ranking by various groups of readers and users, respectively). Note that open access is urgently needed to stimulate innovation by competition in this field, which has long been hampered by monopoly structures. The working capacities of librarians and related information professionals that may be liberated by the end of the subscription business are urgently needed for the structuring, processing, quality assurance, and digital preservation of scientific contents, bibliometric data, and statistical analyses both at scientific institutions and at commercial service providers.

## ACKNOWLEDGMENTS

The author thanks A. Pöschl, M. Pöschl, and M. Weller for inspiration and support. P.J. Crutzen, T. Koop, K.S. Carslaw, R. Sander, W.T. Sturges, A.K. Richter, M. Rasmussen, the scientific communities of ACP and EGU, and the team of Copernicus are gratefully acknowledged for successful and enjoyable collaboration in the development of interactive open access publishing and multi-stage open peer review. Three referees are gratefully acknowledged for stimulating comments in the peer review of this paper.

<sup>45</sup><http://www.springer.com/open+access>

<sup>46</sup>[www.oaspa.org](http://www.oaspa.org)

<sup>47</sup>[www.doaj.org](http://www.doaj.org)

## REFERENCES

- Albarede, F. (2009). AGU announces open peer review experiment. *Eos (Washington DC)* 90, 276.
- Bodenschatz, E., and Pöschl, U. (2008). “Open access and quality assurance,” in *Open Access Challenges and Perspectives – A Handbook*, European Commission and German Commission for UNESCO. Available at: [http://ec.europa.eu/research/science-society/document\\_library/pdf\\_06/open-access-handbook\\_en.pdf](http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-handbook_en.pdf)
- Bollen, J., Van de Sompel, H., Hagerberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., and Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. *PLoS ONE* 4, e4803. doi:10.1371/journal.pone.0004803
- Bornmann, L., and Daniel, H. D. (2010a). Reliability of reviewers' ratings when using public peer review: a case study. *Learn. Publ.* 23, 124–131.
- Bornmann, L., and Daniel, H.-D. (2010b). Do author-suggested reviewers rate submissions more favorably than editor-suggested reviewers? A study on atmospheric chemistry and physics. *PLoS ONE* 5, e13345. doi:10.1371/journal.pone.0013345
- Bornmann, L., Marx, W., Schier, H., Thor, A., and Daniel, H.-D. (2010). From black box to white box at open access journals: predictive validity of manuscript reviewing and editorial decisions at atmospheric chemistry and physics. *Res. Eval.* 19, 105–118.

- Bornmann, L., Neuhaus, C., and Daniel, H.-D. (2011a). The effect of a two-stage publication process on the Journal Impact Factor: a case study on the interactive open access journal atmospheric chemistry and physics. *Scientometrics* 86, 93–97.
- Bornmann, L., Schier, H., Marx, W., and Daniel, H.-D. (2011b). Is interactive open access publishing able to identify high-impact submissions? A study on the predictive validity of atmospheric chemistry and physics by using percentile rank classes. *J. Am. Soc. Inf. Sci. Technol.* 62, 61–71.
- Copernicus. (2011). *A Short History of Interactive Open Access Publishing*. Göttingen: Copernicus Publications.
- David, P. A., and Uhlig, P. F. (2005). “Creating the information commons for e-science: toward institutional policies and guidelines for action,” *Workshop Proceedings*, (Paris: UNESCO).
- Economist Academic Publishing. (2011). *Of Goats and Headaches*. Available at: <http://www.economist.com/node/18744177> [accessed May 26, 2011].
- European Commission and German Commission for UNESCO. (2008). *Open Access Opportunities and Challenges – A Handbook*. Available at: [http://ec.europa.eu/research/science-society/document\\_library/pdf\\_06/open-access-handbook\\_en.pdf](http://ec.europa.eu/research/science-society/document_library/pdf_06/open-access-handbook_en.pdf)
- European Geosciences Union. (2010). *EGU Position Statement on the Status of Discussion Papers Published in EGU Interactive Open Access Journals*. Available at: <http://www.egu.eu/statements/position-statement-on-the-status-of-discussion-papers-published-in-egu-interactive-open-access-journals-4-july-2010.html>
- Golden, M., and Schultz, D. M. (2012). Quantifying the volunteer effort of scientific peer reviewing. *Bull. Am. Meteorol. Soc.* 93, 337–345.
- Lee, C. (2012). Open peer review by a selected-papers network. *Front. Comput. Neurosci.* 6:1. doi:10.3389/fncom.2012.00001
- Max Planck Society. (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. Available at: [http://www.zim.mpg.de/openaccess-berlin/berlin\\_declaration.pdf](http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf)
- Müller, U. (2008). *Peer-Review-Verfahren zur Qualitätssicherung von Open-Access-Zeitschriften – Systematische Klassifikation und empirische Untersuchung*. Ph.D. thesis, Humboldt University, Berlin.
- Pöschl, U. (2004). Interactive journal concept for improved scientific publishing and quality assurance. *Learn. Pub.* 17, 105–113.
- Pöschl, U. (2009a). Interactive open access peer review: the atmospheric chemistry and physics model. *Against Grain* 21, 26–34.
- Pöschl, U. (2009b). Interactive comment on “On the validity of representing hurricanes as Carnot heat engine” by A. M. Makarieva et al. *Atmos. Chem. Phys. Discuss.* 8, S12406–S12411.
- Pöschl, U. (2010a). Interactive open access publishing and public peer review: the effectiveness of transparency and self-regulation in scientific quality assurance. *IFLA J.* 36, 40–46.
- Pöschl, U. (2010b). Interactive open access publishing and peer review: the effectiveness and perspectives of transparency and self-regulation in scientific communication and evaluation. *LIBER Q.* 19, 293–314.
- Pöschl, U. (2010c). *Arne Richter: A multi-talented character who has made a difference in scientific publishing*. EGU General Assembly 2010, Vienna. Available at: [http://www.atmospheric-chemistry-and-physics.net/pr\\_acp\\_poschl\\_arne\\_richter\\_wien2010\\_up31.pdf](http://www.atmospheric-chemistry-and-physics.net/pr_acp_poschl_arne_richter_wien2010_up31.pdf), 2010
- Pöschl, U. (2011). “On the origin and development of interactive open access publishing,” in *A Short History of Interactive Open Access Publishing* (Göttingen: Copernicus Publications).
- Pöschl, U., and Koop, T. (2008). “Interactive open access publishing and collaborative peer review for improved scientific communication and quality assurance,” in *Information Services & Use*, 28 (APE 2008: Academic Publishing in Europe, Quality and Publishing, IOS Press), 105–107. Available at: [http://www.atmospheric-chemistry-and-physics.net/pr\\_acp\\_poeschl\\_koop\\_infoervuse\\_2008\\_intoapub.pdf](http://www.atmospheric-chemistry-and-physics.net/pr_acp_poeschl_koop_infoervuse_2008_intoapub.pdf)
- Priem, J., and Hemminger, B. M. (2012). Decoupling the scholarly journal. *Front. Comput. Neurosci.* 6:19. doi:10.3389/fncom.2012.00019
- Sandewall, E. (2006). Opening of the process. A hybrid system of peer review. *Nature*.
- Sandewall, E. (1997). Publishing and reviewing in the ETAI. *Electron. Trans. Artif. Intell.* 1, 1–12.
- Sandewall, E. (2012). Maintaining live discussion in two-stage open peer review. *Front. Comput. Neurosci.* 6:9. doi:10.3389/fncom.2012.00009
- Schultz, D. M. (2010). Rejection rates for journals publishing in the atmospheric sciences. *Bull. Am. Meteorol. Soc.* 231–243.

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 June 2011; accepted: 21 May 2012; published online: 05 July 2012.

Citation: Pöschl U (2012) Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation. *Front. Comput. Neurosci.* 6:33. doi: 10.3389/fncom.2012.00033  
Copyright © 2012 Pöschl. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



# The evaluation of research papers in the XXI century. The *Open Peer Discussion* system of the *World Economics Association*

Grazia letto-Gillies<sup>1,2\*</sup>

<sup>1</sup> Centre for International Business Studies, London South Bank University, London, UK

<sup>2</sup> International Business, Department of Management, Birkbeck University of London, London, UK

## Edited by:

Diana Deca, TUM, Germany

## Reviewed by:

Antonio Politi, National Research

Council, Italy

Dietrich S. Schwarzkopf, Wellcome

Trust Centre for Neuroimaging at

UCL, UK

Dwight Kravitz, National Institutes

of Health, USA

## \*Correspondence:

Grazia letto-Gillies

e-mail: iettogg@lsbu.ac.uk

The paper starts with a brief discussion of the traditional peer review (TPR) system of research evaluation, its role, and the criticisms levelled at it. An analysis of specific problems in economics leads to a full discussion of the Open Peer Review (OPR) system developed by the World Economics Association (WEA) and the principles behind it. The system is open in the following two respects: (a) disclosure of names of authors and reviewers; and (b) inclusivity of potential reviewers in terms of paradigmatic approaches, country, and community. The paper then discusses the applicability of the same system to other disciplines. In doing so, it stressed the aims of various evaluation systems and the possible pitfalls of rating systems. It also speculates on the future of journal publication.

**Keywords:** research evaluation, academic journals, open peer review, economics, World Economics Association

## INTRODUCTION

The peer review (PR) system of research evaluation used for many decades in journal publications has increasingly come under criticism. In the traditional peer review (TPR)<sup>1</sup> system a small number of reviewers appointed by the editor write reports that form the basis for the decision to publish or not. The names of reviewers are usually kept secret from the authors; the practice on the anonymity of authors varies from journal to journal and indeed between disciplines.

Most criticisms of TPR concentrate on the following<sup>2</sup>: (a) efficiency issues and in particular the high and increasing social costs for the academic community and the length of the publication process (Campanario, 1998a,b; Ginsparg, 2002; Frey and Osterloh, 2007); (b) pressure on authors to accept the suggestions of reviewers—even when they do not agree with them—in order to have their paper published (Frey, 2003); (c) low effectiveness in terms of quality assurance such as the detection of errors or of plagiarism or the weeding out of very poor research (Campanario, 1998a; Bedeian, 2004); and (d) difficulty in identifying groundbreaking research (Horrobin, 1990; Gans and Shepherd, 1994; Campanario, 1995; Gillies, 2008).

There are, nonetheless, many supporters of the TPR system among academics (Ledeberg, 1978; Garfield, 1986; Legendre, 1995). They claim that, though the system does have some faults, it is the best available and one on which there is the widest consensus about its fairness. This view is largely shared by the Report

on the inquiry of the House of Commons Science and Technology Committee (2011).

Most of the problems of the TPR have been known for a long time. It is legitimate to ask ourselves why they have come to the fore now. I suggest that this is due to the following reasons. First, the fact that there has been an increase in assessment in general and researchers are beginning to ask whether it is all necessary and indeed whether this type of culture encourages academic endeavors. Second, there has been an increase in the number of journals and in the number of papers seeking publication—and thus journal space—in response to the widening assessment culture. This proliferation of papers and journals is leading to increasing reviewing work and, indeed, to overload for many reviewers. A third—and in my view most relevant—factor is that the power of digital technologies is making the old system redundant. Essentially, what I am saying is that—whether the commentators realize it or not—our critical attitude to TPR is emerging because *there is a way out*. It is on the basis of this last point—the existence of a way out—that the new system of evaluation—the *Open Peer Discussion* system—in economics was developed as discussed in section “The WEA Evaluation System: Basic Principles and Process.”

However, whether a new system can replace the TPR one largely depends on what we expect from an efficient and effective evaluation system. Most researchers expect it to perform the following functions. (i) Quality assurance for the readers and guidance as regard fields of specialization. (ii) Help in improving the research paper. (iii) Guidance to editors in the allocation of limited journal space.

Regarding (iii), unfortunately the TPR system is known to have led to some perverse allocation: the rejection of papers

<sup>1</sup> Peer review (PR) means review by experts and this includes a variety of systems. For this reason I prefer to distinguish between PR and TPR.

<sup>2</sup> These issues are discussed at greater length in Letto-Gillies (2010). See also Kravitz and Baker (2011) and Birukou et al. (2011) in this issue.

containing fundamental research. Several instances from the history of science modern and past and in several disciplines have come to light (Horrobin, 1990; Gans and Shepherd, 1994; Campanario, 1995). Closer to us, The Guardian (2011) reports that the groundbreaking research of Daniel Shechtman—the 2011 winner of the Nobel Prize for Chemistry—was, at first, rejected by peers and he was asked to leave his research group to which he was, allegedly, bringing disgrace by his theory and findings. Gillies (2008) gives a philosophical reason—based on an application of Kuhn to the research assessment field—of why it should be so. He claims that the TPR system is likely to favor orthodox research, the type of research that operates competently within a well-established and majority paradigm rather than research which is ground-breaking. Yet, the history of science shows that, while the former type of research may be relevant, it is the ground-breaking research that gives science, the economy, and society the best returns in the long run. Sir James Black, the 1988 Nobel Prize winner for medicine, did not mince his words regarding the impact of TPR system on innovative research. In a *Financial Times* (2009) interview he is attributed the following statement: “The anonymous peer review process is the enemy of scientific creativity . . . . Peer reviewers go for orthodoxy . . .”

The next section considers the specific problems of research evaluation in economics and the establishment of the *World Economics Association* (WEA). Section “The WEA Evaluation System: Basic Principles and Process” presents a PR system developed by the WEA and designed to overcome some of these problems. The last section discusses the applicability of the WEA system to other disciplines and emphasizes the desirability to consider the aim of evaluation in developing alternative systems.

### SPECIFIC PROBLEMS IN ECONOMICS. THE WEA

Economics, the dismal science, is also among the most problematic of sciences in terms of research evaluation. The TPR system has been applied in economics as well as—and as long as—in most other sciences—natural or social—and in the humanities. It has drawn a similar amount of criticism.

However, in economics there are also problems that are largely specific to the subject and are additional to the general problems of TPR. Here are some of these specific problems.

First, in economics there is, usually, co-existence of several schools/paradigms contemporaneously. This is one of the features that differentiate the social from the natural sciences according to the philosopher of science Thomas Kuhn<sup>3</sup>. Second, economics and its theories tend to be closely linked to political ideologies and it is this aspect that makes it possible and desirable to have co-existence of several paradigms. Ideology plays a role in the type of issues considered by researchers and economists in general; in how they characterize the operations of the economic system; which methods and empirics—if any—they use to corroborate their theories; and in how they interpret their results<sup>4</sup>. The third problem—common to other

sciences—is that it is possible to earn large amounts of money outside academe as advisers to politicians and consultant for large businesses and institutions. The contact with the real world of business and policy-making may help in the understanding of economic issues and in the development of research; however, it may also affect the objectivity of the researcher. In terms of evaluation of research papers, the referees themselves may be—even unconsciously—biased in favor or against research that is too closely linked to the business or politics they are involved in.

Regarding the first issue—the co-existence of several paradigms—the following alternative paradigms/schools can be identified in economics: Keynesian, Marxist, Sraffian/neo Ricardian, Austrian, institutionalist, and neoclassical. The latter school is the one most closely associated with the following features: supremacy of the market and of its price mechanism as allocator of resources; equilibrium analysis; disregard for uncertainty in economic processes. After the Second World War the Keynesian, neoclassical and Marxist schools were the main paradigms across the western world. To a large extent they coexisted though the Marxist school was always a minority one. It is, however, interesting to note that in those early decades after WWII most economists, whatever the school they belonged to, seemed to accept Keynesian analysis and its policy prescriptions: government intervention to smooth the trade cycle was widely accepted. The Keynesian theory was, in fact, adapted by the neoclassical school to fit in with their equilibrium analysis in the so-called neoclassical synthesis.

In the last 30 years two major changes have occurred in economics. There has been a move toward less pluralism and toward the dominance of the neoclassical school. Moreover, the now prevailing neoclassical school has changed its character compared to its earlier, traditional form. An extreme form of neoclassical economics has now become the dominant paradigm in economics; one with the following features. It: (a) rejects Keynesian analysis and policies; (b) gives the market a supremacy role linked to the belief that unfettered markets can deliver equilibrium and stability; and hence (c) rejects the role of governments in regulating markets<sup>5</sup>. This extreme form of neoclassical economics corresponds, in politics, to the ideology of neoliberalism. As the latter ideology prevailed, so did the supremacy of the neoclassical paradigm in all aspects of economic life and of economics as a discipline; from journal publications to university and school curricula to media analysis and to policy recommendations. Gradually all other paradigms have been marginalized—though not obliterated—and the neo-classical one has become the mainstream paradigm and almost the only one prevailing in terms of policy recommendations.

The TPR system of research evaluation has been one of the key elements in helping the neoclassical system achieve supremacy and making economics almost a single paradigmatic subject. There is an interaction at work: within the TPR system of research

<sup>3</sup>See Gillies (2012) for an account of Kuhn's position on the social versus natural sciences.

<sup>4</sup>The link between economics and ideology does not mean that it is impossible to reach conclusions regarding the validity of results and the corroboration

of theories. Usually, the corroborating or refuting evidence builds up and ideologies can be set aside.

<sup>5</sup>This does not prevent big business and the financial sector asking for government support when they are in trouble.

evaluation a discipline with predominance of a single paradigm will tend to favor publication of papers—particularly in the highest rated journals—within that paradigm<sup>6</sup>. This is largely because most reviewers belong to the mainstream paradigm and are very likely to see negatively papers developed in the context of alternative paradigms. This outcome may not necessarily be the result of deliberate strategies to cut out other paradigms neither of poor judgment: it may, in many cases, be the result of being confronted with something unfamiliar and which, therefore, appears to be not quite right. It must be remembered that many economists currently younger than 50 or so years, may not have been taught any of the alternative paradigms particularly if they have been to very prestigious universities. Moreover, once a paradigm starts prevailing and monopolizing the top journals as well as the allocation of research funds and jobs, more and more young researchers will work within it thus consolidating its supremacy.

Dissatisfaction with this situation and with the dominant economics paradigm—and with the policies it led to—was bound to develop<sup>7</sup>. It has, indeed, increased following the financial crisis of 2008 when the economics profession has come under justified attacks for: (a) having encouraged disastrous economic policies, particularly with regard to financial deregulation; and (b) being at a loss as to what to do once the crisis manifested. Since then the economic situation has worsened and criticisms of the subject and of the profession as a whole continue. It should, however, be noted that the policies of most governments were inspired by main stream type of economics. There were quite a few economists who had been warning against excessive financial deregulation and marketization of economies as they are warning now about promoting deflationary policies in the context of a recession. According to the Keynesian paradigm deflationary policies in the context of low effective demand (for consumption, investment, exports, and government expenditure) lead to lower state revenue and thus they exacerbate the problem of governments' debts. But, alas! these Cassandra voices are not heeded and the relevant papers rarely find their way into prestigious journals or policy circles.

The paradigmatic dominance<sup>8</sup> in the main journals led to concomitant dissatisfaction with the TPR system of research evaluation. The problems were further complicated by the fact that the mainstream paradigm was seen as associated with the dominance of Anglo American economics and economic policies. The *American Economic Association* (circa 17,000 members) and

the old and prestigious *Royal Economic Society* (c. 3300 members) were seen to dominate the type of economics being taught in universities all over the world, the most prestigious journals and—indirectly—the top jobs in finance, politics, and business economics. It also dominated and still dominates the policies of many governments in both developed and developing countries.

It is in this context that the WEA was established. The brain-child of Edward Fullbrook, the WEA was developed with the collaborative effort of a few other people<sup>9</sup> from different parts of the world. All work is done on a voluntary basis by committed people. It was launched on 16th May 2011<sup>10</sup> and within a year it reached a membership of approx 10,000. Membership is free and donations are encouraged.

The WEA aims ([www.worldeconomicsassociation.org](http://www.worldeconomicsassociation.org)) include: plurality of approaches to economics; inclusivity of economists from every part of the world and from every persuasion; commitment to high-level research and to the full utilization of the digital technologies. Its main activities—all online and free to members—are the management of three journals and of conferences. More journals may be developed in the future.

## THE WEA EVALUATION SYSTEM: BASIC PRINCIPLES AND PROCESS

The initiators of the WEA are fully committed to high-level research and to research evaluation. However, they consider that the aims of plurality of approaches to economics and inclusivity could not be achieved within the operation of the TPR system of research evaluation for the reasons explained in section “Specific Problems in Economics. The WEA”. They therefore developed a different system of research evaluation<sup>11</sup> to be used by two of its journals in alternative to the PR system. The journals are: *World Economic Review* (WER) and *Economic Thought: History, Philosophy and Methodology* (ET). The third journal of the Association the *Real World Economics Review* (RWER) has been in operation for several years and is now incorporated into the WEA umbrella. It publishes articles on economic, political, and social issues of wider appeal—and for a wider readership—than the more specialized economics field of the other two journals. The papers are evaluated by the editor of the RWER who publishes what he considers appropriate and after an editing process. The system used in the WER and ET is based on the following principles.

- PR is a very useful system for research evaluation and development. However, the digital technologies have made journal space allocation an irrelevancy. It is therefore possible to decouple the dissemination/publication function<sup>12</sup> from the evaluation and development function of PR.

<sup>6</sup>Gillies (2012) gives a philosophical and mathematical justification of why and how the TPR system of research evaluation in a discipline with prevalence of one paradigm will lead to the highest rating for research in that specific paradigm. Lee (2007) give a statistical analysis of the relationship between scores of journals and adherence to mainstream or minority paradigms. It shows that the highest-rated journals shy away from publishing research papers developed within alternative paradigms.

<sup>7</sup>It has led to the Post-Autistic Economics movement (Fullbrook, 2003) and to the Association of Heterodox Economists (Lee, 2008).

<sup>8</sup>The problems of paradigmatic dominance and power structure are considered in Bachmann (2011) in this issue.

<sup>9</sup>Including the author of this article who contributed, in particular, to the evaluation system for the WEA journals and to the development of its system of online conferences.

<sup>10</sup>Its legal status is of a Community Interest Company.

<sup>11</sup>Most of the points in the WEA alternative system of research evaluation have its origin in Letto-Gillies (2008 and 2010).

<sup>12</sup>See Priem and Hemminger (2012) in this issue.



- The digital technologies are being used extensively by journals' publishers in the publication process. For example in communications between editors, referees, and authors and in copy editing. However, so far, little use has been made of them for the *evaluation process itself*.
- The TPR system is based on the principles of assessment/rating and of exclusion. Because journal space is limited and the ratio of paper submission to acceptances is very high, the editors necessarily look for support and justification for the rejection of many submitted papers. In order to do so, reviewers often look for faults rather than areas which are positive and could be further developed. These critical points do not mean to devalue the work of reviewers<sup>13</sup>—many of whom labor very hard and often come up with helpful suggestions—but only to point out a problem in the system they are caught in: in the end no matter how helpful some of them may want to be, their reports are used to exclude papers from publication in specific journals. But again no blame can be attached to the editors who have to allocate limited space in their journals.
- Research can achieve best results when it is developed as a social activity<sup>14</sup> not necessarily in the sense of two or three people working together on a project, though this is, increasingly, the case in many fields. The social context is seen here as researchers developing their own ideas on the basis of previous research—which is always the case—and benefiting from discussions and interchanges with peers in a constructive environment. The involvement of peers in the evaluation and further development of research is very useful. However, it does not have to be on a confrontational and rating basis. It can take place on the basis of exchange of ideas for the advancement of the specific topic of the paper.
- The involvement of many researchers in the evaluation process is preferable to only 2–3 reviewers because: (a) the large number of reviewers—from an inclusive constituency—is more likely to contain a few who can spot plagiarism, mistakes, data problems; (b) if many people—belonging to several paradigmatic approaches and several countries and communities—read a paper it is more likely that one or two of them spot the originality and value of a paper which is out of the ordinary and may thus appear strange and wrong to most researchers. Thus, one of the major pitfalls of the PR system is less likely to manifest. Moreover, the involvement of many commentators increases the likelihood of researchers belonging to different schools/paradigms contributing. One of the major problems in economics research and publication can, therefore, be avoided.
- Double-sided openness: the names of the author(s) and those of reviewers are revealed. The attribution of comments to a specific paper encourages commentators to come forward with their views knowing that they are posted with their names. Attribution may, therefore, eliminate reticence in putting forward very original comments. Attribution may also encourage commentators to consider carefully their critical arguments and make sure that they are not inspired mainly by adherence to a specific paradigm and ideology.
- A common worry about open posting (where the names of authors and commentators are disclosed) is that commentators feel embarrassed to be critical. However, it is worth pointing out that: (a) reviewers of books—where a doubly open system is used as a matter of normal academic activity—are often quite critical; (b) moreover, if the process is online, commentators, and authors may be in very distant parts of the world and do not know each other; and (c) if the system is less confrontational than the TPR system this is no bad thing: a critically positive system is more likely to lead to the advancement of research.
- Post-publication evaluation is as important for the advancement of research as pre-publication one. The life of a paper does not end with publication; hopefully that is only its beginning. Other researchers will read the paper for years to come; the continuing readership success of the paper through time is evidence of its relevance. Some readers may develop further research of their own after reading an article and their research may lead to new publications in their own name. However, others may have points to make about it which do not amount to the development of a full research project or paper but that can, nonetheless, be relevant and useful for the further advancement of the field. A *post-publication commentary* as a standard feature of journals allows these people to have their comments published—at the discretion of the editors—with attribution.

The above principles inform the WEA system of *Open Peer Discussion (OPD)* whose actual process is the following<sup>15</sup>.

1. Papers submitted to the journal are first vetted by the editors. Those that meet minimum standards of professional quality are posted with the name of the author on the journal's *Discussion Forum (DF)*. Each (DF) remains open for eight weeks from the posting of the paper. All members of the WEA have access to the DF and can actively participate in it.
2. Comments on the posted paper are invited from the membership as well as solicited by the editors from experts in the field. Names of possible commentators may also be suggested by the authors. The comments are screened by the editors and then posted with the name of the commentator unless anonymity is requested. The authors can respond to the comments and their response will be posted with attribution.
3. Once the DF is closed the editors reach their decision on whether to publish the paper and—if accepted—the author is invited to review the paper for publication. Selected important reviews will be published at the end of the paper with prior agreement from the commentators.

<sup>13</sup>This author has been associate editor of an academic journal—*Transnational Corporations*—and has been involved for many years in TPR activity as a reviewer and as an author. In the latter activity she has received many helpful comments and some lousy ones.

<sup>14</sup>Lee (2012)'s approach in this issue is also based on the principle that open, attributed reviews by many researchers can help to further develop a specific paper. However, the recommended system is different from the one presented here.

<sup>15</sup>Some of the following elements of OPR are applied also to the WEA conferences in which the text-based discussion takes place online over a four weeks period.

4. A *Post-Publication Commentary* section is open on the journal. Post-publication comments are sent to the editors who will decide whether to post them or not.

### FROM THE SPECIFIC TO THE GENERAL. A VISION FOR FUTURE EVALUATION SYSTEMS

The previous two sections presented an application of an Open Peer Review (OPR) system to the case of economics. To what extent can this specific case be generalized to apply to other disciplines? In order to answer this question let us consider what elements are necessary for the system to operate and whether those elements can be had in other disciplines.

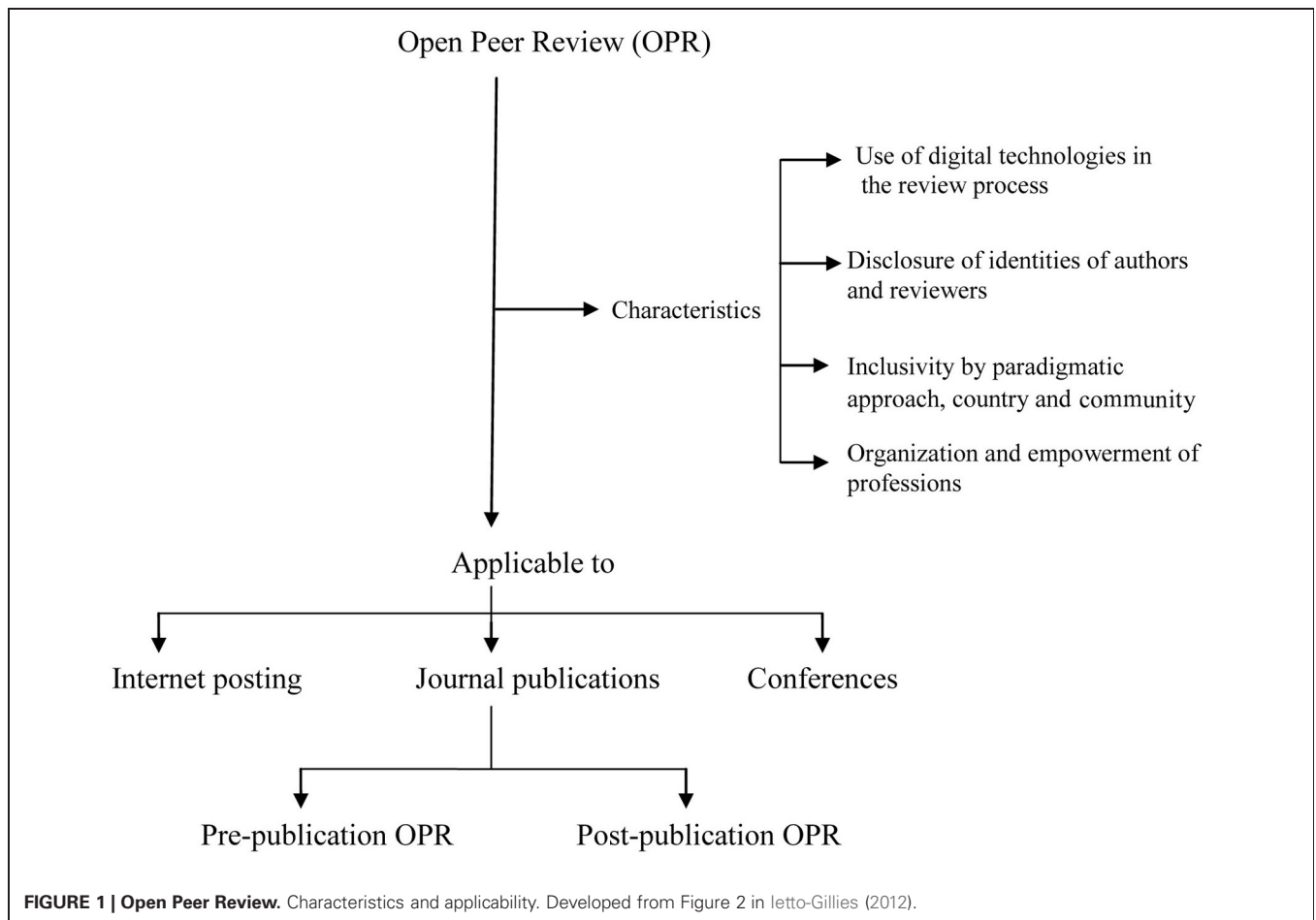
The system is open in two respects: (1) because there is attribution of authorship for both authors and reviewers; and (2) because there is inclusivity of potential reviewers in terms of paradigmatic approaches, countries, and communities. Point (2) requires (a) the use of digital technologies; and (b) the full involvement and empowerment of the research community in any specific field. In order to realize point (b) it is necessary to be inclusive and thus to reach a large number of diverse potential reviewers. This is now possible via the digital technologies which, therefore, enter into the very process of research evaluation rather than contribute only via the digitalization of administrative functions. Point (1) is more likely to lead to reviewing that

is: more carefully thought through; less likely to be biased and more likely to lead to comments that make positive points towards the development of research. Point (2b) raises the probability of the reviewers being able to spot errors, fraud or ground breaking research.

**Figure 1** illustrates the elements on which the OPR system is based as well as its possible applications: to journal publications, to internet posting and to online conferences. These requirements can be had for all or most disciplines and therefore I see the possibility of applying the OPR system discussed above to fields other than economics. If disciplines are very large in terms of members—as is, indeed, economics—it may be necessary to classify the members by fields of specific interest. Participation to the OPR process would then be limited to researchers that specialize in the field of the paper to be reviewed.

It should, however, be noted that—quite independently of the discipline—just the reaching of many diverse researchers is no guarantee of having a large and diverse contribution to OPR. In fact early experiences<sup>16</sup>—including that of the WEA—point to the fact that researchers are timid about exposing themselves as

<sup>16</sup>See Pöschl (2012) in this issue. The Science and Technology Committee (2011) report the successful case of the *British Medical Journal* and the unsuccessful one of *Nature* (p. 26, para 23).



**FIGURE 1 | Open Peer Review.** Characteristics and applicability. Developed from Figure 2 in letto-Gillies (2012).

authors or reviewers. This is not surprising given the culture of secrecy in PR to which we are all used. It will take time, but I believe that a change in culture is possible. Meanwhile editors may want to consider starting the system with a mixture of anonymous and attributed reviews and may, generally, be prepared to be flexible in terms of allowing anonymity in special cases.

How do we progress from current positions toward the implementation of OPR systems? I see various possible routes. First, a bottom up approach; this is the one used in the WEA case and the only route of which I have direct experience. Volunteers within the discipline work toward the establishment of a new association with the specific objectives of organizing online journals and conferences. This initial process involves a considerable amount of work, commitment, and goodwill. Alongside efforts to increase the membership and promote the activities, there will be efforts to set up the activities such as: appointing editorial boards and editors and producing tight guidelines for both conferences and journals. A second route would be to start from existing associations and propose to members OPR-led activities. A third route is to start from existing journals and encourage the readership to opt for OPR processes and also to participate in these processes as authors and reviewers.

Though, in this writer's view, the system is not discipline-specific it does have boundaries in respect to other elements. First, in terms of aims. The main aim of the system presented in this paper is to contribute to the development of research. However, the reviewing process may be developed—with the aid of digital technologies—to meet other aims. For example, to help readers—who may or may not themselves be researchers—to find their way through large number of published works. There are several initiatives in this direction such as the Faculty of 1000<sup>17</sup> for the biological sciences. Similar aims are behind the development of quick, snappy ratings of papers, a practice that is spreading fast. Personally I am not in favor of these types of rating evaluations: they stress the competitive side of research rather than the collaborative and social nature of research and, moreover, they lend themselves to abuse and to possible misinterpretations by the readers. The digital technologies offer us many possibilities for rating purposes and we are in danger of developing more and more rating systems just because the technology allows us to. In other words we are in danger of being technology-led rather than aim-led with the technology being used to meet specific aims.

Whether we are in favor of rating or not, in my view the key question to ask ourselves is: what aim do we want to achieve by rating? How can the technology help us to achieve those aims?

The possible developments in PR systems discussed in this paper and, indeed, in this journal issue speak for a future evaluation system different from the current one. Moreover, if we consider the combination of open access (OA) systems in the field of dissemination and of OPR system in the evaluation field<sup>18</sup> we may be led to speculations about the future of publication via journals. In the discourse on evaluation—including the OPR system presented above—the starting point is publication and how to develop an alternative system of evaluating papers pre- and post-publication.

However, let us put “evaluation with the aim of development” center stage and let us assume that some system of OPR becomes widespread. We can then speculate whether in such a future we shall still need journals—be they in electronic or paper version. We shall still need “editors” to manage the evaluation function. However, once the paper has been openly reviewed and revised and receives the approval of the editors, do we need it to be bundled up with other papers and be published as part of a journals issue? What are the benefits of such bundling and publication process? Could it not just be posted on an OA repository labeled something like “evaluated and revised papers”? It might still be possible to have comments on these finalized papers and even have the authors write “Addendums” to their papers if they later want to make further developments to it. Regarding the bundling together in a journal issue, might there still be scope for this practice but in terms of bundling up by topic? Might readers find it more useful to have papers bundled up by topic rather than by the date at which various papers happen to be ready for publication?

I do not have answers to many of these questions. The field and the issues are evolving. The only thing I am sure of is that the future of dissemination and evaluation of research will look very different from the present.

## ACKNOWLEDGMENTS

I am grateful to Diana Deca, Nikolaus Kriegeskorte, and three anonymous referees of this journal for their helpful comments.

<sup>18</sup>The relationship between OA and OPR is discussed further in Ietto-Gillies (2012).

<sup>17</sup>[www.f1000.com](http://www.f1000.com)

## REFERENCES

- Bachmann, T. (2011). Fair and open evaluation may call for temporarily hidden authorship, caution when counting the votes, and transparency of the full pre-publication procedure. *Front. Comput. Neurosci.* 5:61. doi: 10.3389/fncom.2011.00061
- Bedeian, A. G. (2004). Peer review and the social construction of knowledge in the management discipline. *Acad. Manag. Learn. Educ.* 3, 198–216.
- Birukou, A., Wakeling, J. R., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K., Osman, N., Ragone, A., Sierra, C., and Wassef, A. (2011). Alternatives to peer review: novel approaches for research evaluation. *Front. Comput. Neurosci.* 5:56. doi: 10.3389/fncom.2011.00056
- Campanario, J. M. (1995). Commentary: on influential books and journal articles initially rejected because of negative referees' evaluations. *Sci. Commun.* 16, 304–325.
- Campanario, J. M. (1998a). Peer review for journals as it stands today—Part 1. *Sci. Commun.* 19, 181–211.
- Campanario, J. M. (1998b). Peer review for journals as it stands today—Part 2. *Sci. Commun.* 19, 277–306.
- Frey, B. S. (2003). Publishing as prostitution? Choosing between one's own ideas and academic failure. *Public Choice* 116, 205–223.
- Frey, B. S., and Osterloh, M. (2007). *Evaluations: Hidden Costs, Questionable Benefits, and Superior Alternatives*. Zurich: IEW Working Paper Series No. 302, February.
- Fullbrook, E. (2003). *The Crisis in Economics. The Post-Autistic Economics Movement*. London, UK: Routledge.
- Gans, J. S., and Shepherd, G. B. (1994). How the mighty have fallen: rejected classic articles by leading economists. *J. Econ. Perspect.* 8, 165–179.
- Garfield, E. (1986). Refereeing and peer review. Part 2. The research on refereeing and alternatives to the

- present system. *Curr. Contents* 11, 3–12.
- Gillies, D. (2008). *How Should Research Be Organised?* London, UK: College Publications.
- Gillies, D. (2012). “Economics and research assessment systems,” in *Economic Thought. History, Philosophy and Methodology*, Vol. 1, 23–47.
- Ginsparg, P. (2002). “Can Peer Review be better Focused?” Available online at <http://people.ccmr.cornell.edu/ginsparg>
- Horrobin, D. F. (1990). The philosophical basis of peer review and the suppression of innovation. *J. Am. Med. Assoc.* 263, 1438–1441.
- House of Commons Science and Technology Committee, (2011). ‘Peer Review in Scientific Publications’ Eighth Report of Session 2-010-12, Vol 1, HC 856, July, London, UK: The Stationary Office Ltd. <http://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/peer-review/>
- Ietto-Gillies, G. (2008). A XXI-century alternative to XX-century peer review *Real-World Economics Review*, 45, 10–22, March. [www.paecon.net/PAERReview/issue45/IettoGillies45](http://www.paecon.net/PAERReview/issue45/IettoGillies45)
- Ietto-Gillies, G. (2010). “A XXI alternative to XX century Peer Review,” in *Production, Distribution and Trade: Alternative Perspectives. Essays in Honour of Sergio Parrinello*, eds A. Birolo, D. Foley and H. D. Kurz, (London, UK: Routledge), 333–348.
- Ietto-Gillies, G. (2012). Open peer review, open access and a house of commons report, *Real World Economics Review*, issue n. 60, 20 June, 74–91. <http://rwer.wordpress.net/PAERReview/issue60/IettoGillies60.pdf>
- Kravitz, D. J., and Baker, C. I. (2011). Toward a new model of scientific publishing: discussion and a proposal. *Front. Comput. Neurosci.* 5:55. doi: 10.3389/fncom.2011.00055
- Lee, F. (2007). The research assessment exercise, the state and the dominance of mainstream economics in british universities. *Cambridge J. Econ.* 31, 309–325.
- Lee, F. (2008). “Heterodox economics,” in *The New Palgrave Dictionary of Economics*, 2nd Edn. eds S. N. Durlauf and L. E. Blume (Palgrave Macmillan). <http://www.dictionaryofeconomics.com/article>
- Lee, C. (2012). Open peer review by a selected-papers network. *Front. Comput. Neurosci.* 6:1. doi: 10.3389/fncom.2012.00001
- Legendre, A. M. (1995). Peer review of manuscripts for biomedical journals. *J. Am. Vet. Med. Assoc.* 207, 36–38.
- Pöschl, U. (2012). Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation. *Front. Comput. Neurosci.* 6:33. doi: 10.3389/fncom.2012.00033
- Priem, J., and Hemminger, B. M. (2012). Decoupling the scholarly journal. *Front. Comput. Neurosci.* 6:19. doi: 10.3389/fncom.2012.00019
- The Financial Times. (2009). “An acute talent for innovation,” February 2nd.
- The Guardian. (2011). Nobel vindication for once-ridiculed researcher. Oct 6th, 21.

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 05 November 2011; accepted: 06 July 2012; published online: 07 August 2012.

Citation: Ietto-Gillies G (2012) The evaluation of research papers in the XXI century. The Open Peer Discussion system of the World Economics Association. *Front. Comput. Neurosci.* 6:54. doi: 10.3389/fncom.2012.00054

Copyright © 2012 Ietto-Gillies. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



# Post-publication peer review: opening up scientific conversation

Jane Hunter\*

Faculty of 1000, Ltd., London, UK

\*Correspondence: jane.hunter@f1000.com

**Edited by:**

Diana Deca, Technical University Munich, Germany

**Reviewed by:**

Diana Deca, Technical University Munich, Germany

## CONVENTIONAL PEER REVIEW: RIGHTS AND WRONGS

Peer review is broken. We have all heard that phrase many times in recent years. It's become a truism, a shorthand complaint about the *status quo* that rarely extends into a proposal for change. And even those who do not believe standard peer review is beyond repair acknowledge that there are problems; everyone can see the cracks.

So what's wrong? From an author's point of view, a lot. Peer review is slow; it delays publication. It's almost always secret; authors do not know who is reviewing their work – perhaps an ally but, equally, perhaps a competitor. It can block ingenuity; think of the classic case of Lynn Margulis and the 15 or so journals that rejected her ground-breaking article “On the origin of mitosing cells” (Sagan, 1967) before it was finally accepted by *The Journal of Theoretical Biology*. And there's a lot wrong for reviewers too: what proportion of referee reports are second, third, or even fourth round reviews? A referee's hard work may be contributing nothing new to an author who would rather take his or her chances with another journal than do the extra work suggested by reviewers for journals one to three.

Does conventional peer review work for publishers? Well, yes and no. Yes, at top-flight journals like *Nature* or *NEJM* peer review is a gate keeper that helps guarantee publication of only the most interesting articles, and yes, in theory, it helps guard against the publication of flawed work, but it's expensive – even though reviewers work for free – and it's time-consuming. *Nature* or *NEJM* review thousands of papers each year that would not make it into their journals; for third-, fourth-, or fifth-tier journals, somewhere further down the inevitable cascade,

referees will often be doing work that has been done already on an article that was written months ago.

If standard peer review is intended to help ensure that an article is good enough to be published, is it working? And in this context, what does “good enough” even mean? Since most papers will eventually be published, cascading until they find a journal, that means that most papers are good enough for someone and peer review's supposed qualitative gatekeeper role is not supportable. The impact of peer review on the publication of an article is not so much a question of yes or no, it's more likely to be a question of when and where.

Yet even acknowledging the flaws, redundancies, and costs of the conventional peer review system, it is clear that we need peer review. The more specialized science becomes the more we must rely on experts to help us navigate the multiplicity of subject areas we are not expert in ourselves. Peer reviewers are those experts and we depend on the refereeing process to protect us from sloppy work and invalid conclusions.

So peer review is important but the way it happens is problematic

At F1000, we believe that most of the weaknesses of standard peer review can be linked to two core issues, first that it is conducted pre-publication and second that it is secret. Pre-publication peer review allows journals and reviewers to delay, filter, and interrupt the essential conversation of science, and secrecy makes these problems impossible to resolve.

## POST-PUBLICATION PEER REVIEW: TWO MODELS FROM FACULTY OF 1000

A little background: faculty of 1000 began in 2002 with a post-publication review service called F1000 Biology. Its remit was (and still

is) to work with named experts to identify and recommend the most interesting papers published across 24 different subject areas in biology. In 2006 F1000 Medicine joined it – with the same aim, more experts and coverage of 20 medical specialties. We merged the two services in 2010, and biology and medicine are now both covered at F1000.com.

Since then, we have launched F1000 Posters, an open access repository for posters and presentations – again in biology and medicine – and we are now in the early stages of launching our new open access, post-publication peer review journal, *F1000 Research*.

Faculty of 1000 practices two forms of post-publication peer review: primary, open refereeing of articles after they are published in *F1000 Research*, and secondary peer review of the best already-refereed articles, published in any biology or medicine journal, at F1000.com. Both are illustrations of Clay Shirky's “publish then filter” model (Shirky, 2008) and each adds value to scientific discourse in its own way.

I will describe our secondary post-publication review process first.

## SECONDARY POST-PUBLICATION PEER REVIEW

The F1000 article recommendation service applies a layer of positive filtering on top of traditionally peer reviewed literature; we review already-published biology and medicine in order to identify and promote the best work. Our 10,000 named Faculty Members and their Associates select articles that impress them, regardless of source, and write brief recommendations explaining what makes the work significant and putting the science in perspective. These recommendations and comments, along with links to the original articles, are published on F1000.com.



Why is this a useful thing to do? It's useful because the vast volume of material published each year (or each day) makes it difficult for researchers to stay up to date with their own specialized fields, let alone with peripheral fields – all those other subject areas they should be keeping an eye on. Sure, you can search for articles and find, more or less, what you are looking for, but it's helpful to have access to expert opinion for timely guidance on what's especially significant and why. The fact that F1000's reviewers are named puts their opinions in perspective. No one has ever suggested that our F1000 Faculty Members should conduct this form of post-publication review anonymously.

## PRIMARY POST-PUBLICATION PEER REVIEW

*F1000 Research*, F1000's new primary open access publishing program in biology and medicine, publishes immediately, and offers fully open, post-publication peer review. We published our first articles in mid-July and are planning for a full launch at the end of this year.

Articles submitted to *F1000 Research* are first processed through an in-house sanity check and then, assuming they pass, published immediately. Post-publication they are subjected to formal peer review. Referees' reports are published on the site and all referees are named.

The most important task for our referees is to tell us immediately whether or not an article is good science. We do not need to know if it's exciting, or novel, or groundbreaking, we simply want to know that it's valid; that it's sensible work, carefully done. We expect the vast majority of submissions to be approved as good science. If it is good science, an article will be marked as such. If it's not, or if it's good science but the referee has reservations, we require that the referee add a report describing the problems and – if applicable – suggesting improvements. We encourage, but do not require, referees to add reports to articles they have approved as good science.

Authors have the opportunity to respond to a referee's comments and are encouraged to update their articles and

publish revised versions on the site. All versions are separately citable. All articles and all versions are clearly marked with their referee status and articles that have not yet been refereed are labeled as "Awaiting Review."

The strengths of this model are that it's fast, all good science can be published immediately and become part of the record to the benefit of scientists and others worldwide; it's fair, publication cannot be blocked or slowed by the refereeing process; and it's open, and openness discourages bias.

We do not see many weaknesses or risks with this model ourselves – standard peer review has few fans and is overdue for change – but then you might expect us to say that. We do understand though that there are concerns. These include:

- *Is there a risk that F1000 Research will publish junk?*: No, there is not. It will publish good science and let the community decide what the ultimate value of a specific piece of work is. As an aside, we expect that less junk – however one might define that term in science – will be submitted to *F1000 Research* than to conventional journals because few people will want to see a severely negative review of their work become part of the public record. Because *F1000 Research* will publish immediately then review openly, sloppy work will be publicly described as such.
- *OK, if not junk then uninteresting science*: Maybe, maybe not. Uninteresting science is still science, and we believe it should be published. There is a reason for top-line journals to sharply restrict what they publish, that's how they create and maintain their identities and Impact Factors, but it's hard to argue that such restrictions on scientific discourse are, overall, a good thing. We believe they are not. Valid science should be published.
- *No reviewer will want to be openly negative about another scientist's work*: Having now published our first articles we are seeing in real time that this is not the case. Referees are happy to criticize and authors are happy to be able

to respond, to present their case. And because everything is happening in the open, interested scientists can, for the first time, read the back-and-forth and make up their own minds.

*F1000 Research's* version of "publish then filter" is an innovation in life-science publishing and no doubt additional concerns will arise as we fine-tune our model. However, it's clear to us that the research community as a whole is more than ready to contemplate and, we believe, support real change. Complaints about conventional, pre-publication, closed peer review systems are mounting and the risks associated with our "publish first/referee openly later" system seem relatively trivial when compared with the increasing expense and frustration associated with the *status quo*.

We were the inventors of and original advocates for open access. We created Biomed Central, helped set up PubMed Central, and fought the publishing establishment for years to prove that open access can work, that it can be a profitable alternative to standard subscription models. *F1000 Research* and its novel publishing model take openness to the next level. Open access removes barriers for readers. Open, post-publication refereeing removes barriers for readers and authors alike, and it refocuses the role of peer review from, at its worst, a behind-the-scenes variety of censorship to, at its best, the process of expert criticism and advice that has always been its core and upon which the progress of science depends.

## REFERENCES

- Sagan, L. (1967). On the origin of mitosing cells. *J. Theor. Biol.* 14, 225–193.
- Shirky, C. (2008). *Here Comes Everybody: The Power of Organizing Without Organizations*. Penguin Press.
- Received: 03 August 2012; accepted: 07 August 2012; published online: 30 August 2012.
- Citation: Hunter J (2012) Post-publication peer review: opening up scientific conversation. *Front. Comput. Neurosci.* 6:63. doi: 10.3389/fncom.2012.00063
- Copyright © 2012 Hunter. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.



# Designing next-generation platforms for evaluating scientific output: what scientists can learn from the social web

Tal Yarkoni \*

*Institute of Cognitive Science, University of Colorado Boulder, Boulder, CO, USA*

**Edited by:**

Nikolaus Kriegeskorte, Medical  
Research Council Cognition and  
Brain Sciences Unit, UK

**Reviewed by:**

Nikolaus Kriegeskorte, Medical  
Research Council Cognition and  
Brain Sciences Unit, UK  
Chris I. Baker, National Institutes of  
Health, USA

**\*Correspondence:**

Tal Yarkoni, Institute of Cognitive  
Science, University of Colorado  
Boulder, UCB 345, Boulder,  
CO 80309, USA.  
e-mail: tal.yarkoni@colorado.edu

Traditional pre-publication peer review of scientific output is a slow, inefficient, and unreliable process. Efforts to replace or supplement traditional evaluation models with open evaluation platforms that leverage advances in information technology are slowly gaining traction, but remain in the early stages of design and implementation. Here I discuss a number of considerations relevant to the development of such platforms. I focus particular attention on three core elements that next-generation evaluation platforms should strive to emphasize, including (1) open and transparent access to accumulated evaluation data, (2) personalized and highly customizable performance metrics, and (3) appropriate short-term incentivization of the userbase. Because all of these elements have already been successfully implemented on a large scale in hundreds of existing social web applications, I argue that development of new scientific evaluation platforms should proceed largely by adapting existing techniques rather than engineering entirely new evaluation mechanisms. Successful implementation of open evaluation platforms has the potential to substantially advance both the pace and the quality of scientific publication and evaluation, and the scientific community has a vested interest in shifting toward such models as soon as possible.

**Keywords:** data sharing, open access, peer review, publishing, scientific evaluation

Archimedes is widely considered one of the greatest mathematicians and scientists of antiquity. Yet he lived during a period of history (the third century BC) not known for meticulous record keeping, and our appreciation of his seminal contributions consequently depends largely on good fortune. Because of his correspondence with the scholars Conon and Dositheus at the library of Alexandria, we now know of his seminal contributions to geometry and mechanics—work that formed the basis of numerous engineering advances and mathematical discoveries in subsequent centuries (Heath, 1897; Chondros, 2010a,b). But any numbers of slight deviations in the course of history—say, a crucial letter lost at sea, or a librarian's decision to reuse one of Archimedes' palimpsests—could have resulted in the permanent loss of his seminal works (and indeed, a number have never been recovered). In Archimedes' time, and through most of modern human history, the rate of scientific and technological progress depended not just on who discovered what, but also on how good people were at preserving knowledge of what they discovered for future generations. And since record keeping was a difficult business that involved allocation of limited resources, progress also depended heavily on the scholarly community's collective ability to accurately determine *which* work was worth keeping around for posterity.

Modern technology has now solved the problem of preservation; contemporary scientists can rest assured that virtually every scientific article published today will exist in digital form

in perpetuity. One might intuitively expect that this alleviation of the preservation bottleneck would also eliminate the selection problem; after all, if it costs virtually nothing to publish and preserve, why bother to suppress a scientific manuscript that could be useful to someone else down the line, however improbable the odds? Yet in many respects, the scientific community still behaves as though record keeping were a difficult enterprise and paper a scarce commodity. We spend months waiting to hear back from reviewers at journals with 90% rejection rates, anguishing over the prospect that our work might not see the light of day, even though we could disseminate our manuscript to the whole world at any moment via the web. We rely heavily on a select few individuals to pass judgment on our work, even though dozens or hundreds of other researchers are likely to form an informed opinion of its merits within days of official publication. And while we wait for the reviews to come in, we silently fret over the possibility that we might be “scooped” by someone else, even though all it takes to establish scientific precedence is one timestamped upload in a preprint repository.

The continued reliance on an anachronistic publication and evaluation model is striking given the widespread awareness of its many limitations (Smith, 2006; Casati et al., 2007; Jefferson et al., 2007; Young et al., 2008; Ioannidis et al., 2010). Many scientists seem all too happy to move away from the current publishing model and adopt an alternative model that emphasizes open access and “crowdsourced” evaluation. But progress toward

such a goal has been relatively slow. While preprint servers such as arXiv.org have attained near universal usage in some disciplines, such platforms provide few if any tools for evaluation of manuscripts. Conversely, the few platforms that do allow users to evaluate manuscripts post-publication (e.g., the Public Library of Science's platform; <http://plos.org>) have a restricted scope and limited userbase (e.g., analysis of publicly available usage statistics indicate that as of this writing, PLoS articles have received an average of 0.06 ratings and 0.15 comments; <http://www.plosone.org/static/almInfo.action>).

Understandably, norms take time to change; what's surprising is perhaps not that scientists still rely on publishing and evaluation models developed centuries ago, but that they do so in the face of available alternatives. While the scientific community has been slow to embrace emerging information technology, that technology has itself evolved very quickly, and now supports tens of thousands of websites featuring a prominent social component—what has come to be known as the *social web*. In many respects, the challenges faced by popular social web applications—spanning everything from Amazon to Netflix to reddit to Last.fm—closely resemble those involved in evaluating scientific work: How can we combine disparate ratings from people with very different backgrounds and interests into a single summary of an item's quality? How do we motivate users to engage with the platform and contribute their evaluations? What steps should we take to prevent people from gaming the system? And can we provide customized evaluations tailored to individual users rather than the userbase as a whole?

In the rest of this paper, I discuss a number of principles that should guide the implementation of novel platforms for evaluating scientific work. The overarching argument is that many of the problems scientists face have already been successfully addressed by social web applications, and developing next-generation platforms for scientific evaluations should be more a matter of adapting the best currently used approaches than of innovating entirely new ones (cf. Neylon and Wu, 2009; Priem and Hemminger, 2010). Indeed, virtually all of the suggestions I will make have, in one form or another, already been successfully implemented somewhere on the web—often in a great many places.

I begin by briefly reviewing the limitations of the current publishing and evaluation model. I argue that since a transition away from this model is inevitable, and is already in progress, it behooves us to give serious thought to the kinds of platforms we would like to see built in the near future—and increase our efforts to implement such platforms. I then spend the bulk of the article focusing on three general principles we should strive to realize: openness and transparency, customizability and personalization, and appropriate incentivization. Finally, I conclude with a consideration of some potential criticisms and concerns associated with the prospect of a wholesale change in the way the scientific community evaluates research output.

## LIMITATIONS OF CURRENT PRACTICE

Although the focus of the present article and others in this collection is on constructive ideas for new scientific evaluation platforms rather than on critiques of existing models, a brief review of some major limitations of current evaluation practices will

provide a useful backdrop for subsequent discussion of alternative approaches. These limitations include the following.

### SLOWNESS AND INEFFICIENCY

Most articles that eventually get published in peer-reviewed journals go through several cycles of revision and re-review—often at different journals. Typically, months or years elapse between the initial submission and official publication of a manuscript (Ray, 2000; Ellison, 2002; Hall and Wilcox, 2007; Kravitz and Baker, 2011). Most of that time is spent passively waiting rather than actively revising or reviewing; authors have to wait for editors, editors have to wait for the slowest reviewer, and when a paper is rejected, everyone has to wait for the authors to revise and resubmit the manuscript to a different journal. There's no principled justification for such delays and inefficiencies; they simply fall out of current publishing models, with many journals having to reject the vast majority of submissions received in order to preserve a reputation for quality and selectivity. Improving the speed and efficiency of the review process could potentially have a dramatic impact on the rate of scientific progress.

### OPACITY

Because the peer review process is typically conducted behind closed doors, most reviews leave no cumulative record for other scientists to peruse, and allow no independent evaluation of the reviews or reviewers themselves. The problem with lack of transparency is that the quality of reviews is highly variable, frequently leading to rejection of articles for spurious reasons (see below). Unfortunately, under the current model, consumers have no way to evaluate the process that led up to a final decision, or to review any of the interactions between authors, reviewers, and editors. This opacity increases the likelihood of incorrect judgments about a paper's merits, and runs completely counter to the cumulative and open nature of the scientific enterprise. If we don't know who said what about a manuscript and how the manuscript's authors responded, we run a high risk of overlooking or repeating potentially important mistakes.

### LOW RELIABILITY

Current evaluation practices might be defensible if there were empirical evidence that such practices achieve their goals; but formal studies consistently suggest that conventional pre-publication peer review is of limited utility in establishing the quality of manuscripts (though it is undeniably better than no peer review at all). A recent random-effects meta-analysis of 48 studies, comprising 19,443 manuscripts, estimated an inter-rater intra-class correlation of only 0.34 (Bornmann et al., 2010). Since most articles are evaluated by only two or three reviewers prior to publication, and editorial decisions typically follow those of reviewers, it follows that many decisions to accept or reject a manuscript are not appreciably better than chance. This point is corroborated by the grossly uneven distribution of citation rates for articles published in top journals such as *Nature* and *Science* (which explicitly select articles on the basis of perceived impact): a minority of articles typically account for the vast majority of citations, and a sizeable proportion of published articles receive few or no citations (Seglen, 1997; Dong et al., 2005; Mayor, 2010).

While citation counts are not a direct measure of paper quality, there is little reason to suppose that journal impact factor predicts other metrics or expert judgments any better. To the contrary, retrospective evaluations have found modest or no correlations between journal impact factor and expert ratings of impact or quality (Bath et al., 1998; West and McIlwaine, 2002; Maier, 2006; Sutherland et al., 2011). Such findings imply that the heavy emphasis scientists often place on “high-impact” publications when evaluating other researchers’ work is likely to be misplaced.

### LACK OF INCENTIVES

Reviewing scientific manuscripts is time-consuming and effortful. Unfortunately, peer reviewers have relatively little incentive to do a good job. Outside of a sense of duty to one’s profession and peers, and perhaps a pragmatic desire to curry favor with editors, scientists have little to gain by volunteering their time as reviewers, let alone by turning in high-quality reviews on time (Mahoney, 1977; Hojat et al., 2003). Indeed, in some cases, reviewers may even have incentives to write *bad* reviews—for instance, when a researcher is asked to evaluate a competitor’s manuscript. There’s no doubt that the vast majority of scientists will do the right thing in such cases; but it surely seems like bad policy to rely on a system that depends almost entirely on communal goodwill. An ideal evaluation model would directly incentivize the behaviors that maximize the success of the scientific enterprise as a whole, and conversely, would actively deter those that threaten the quality or efficiency of that enterprise.

### A TRANSITION IS INEVITABLE

The limitations reviewed above exist for good reasons, of course. But those reasons are almost entirely historical. When papers were published exclusively in print and scientific communication took place via the postal service, it made sense to restrict publication to a minority of papers that passed some perceived litmus test for quality. But such constraints don’t apply in an age of electronic communication, open access repositories, and collaborative filtering algorithms. Now that the marginal cost of replicating and disseminating manuscripts has dropped to essentially nothing, it makes little sense to artificially restrict the availability or flow of scientific information. There’s a continued need for quality control, of course; but that can be achieved using “soft” filtering approaches that dynamically emphasize or deemphasize information *ad hoc*. It doesn’t require destructive approaches that permanently remove a large part of relevant data from the record. If Archimedes in his day had had the option of instantly depositing his work in arXiv, it’s doubtful that anyone today would accuse him of wasting a few bytes. It’s relatively easy to ignore information we don’t need, but not so easy to recreate information that no longer exists.

One might argue that flooding the scientific literature with papers that have received little or no prior scrutiny would result in information overload and make it impossible to separate good research from bad. But whatever the merit of this argument (and I argue below that it has little), it seems clear at this point that the ship has already sailed. With a modest amount of persistence, scientists can now place virtually any manuscript in a peer-reviewed

journal somewhere (Chew, 1991; Ray, 2000; Hall and Wilcox, 2007)—and often in well-respected venues. For instance, PLoS ONE, the world’s largest journal, published over 7000 articles in 2010, spanning nearly all domains of science, and accepted approximately 70% of all submissions (<http://www.plosone.org/static/review.action>). This model appears so financially successful that Nature Publishing Group and SAGE have both recently launched their own competing open-access, broad-scope journals (Scientific Reports and SAGE Open). To put it bluntly, between megajournals like PLoS ONE and thousands of specialized second- and third-tier journals, we already *are* publishing virtually everything. But we’re doing it very slowly and inefficiently. So the real question is no longer whether or not the scientific community should transition to an open publishing model (Harnad, 1999; Shadbolt et al., 2006); it’s how to handle the inevitable flood of information most efficiently and productively. Our current approach is to rely on heuristics of dubious value—e.g., journal impact factors. But there are far better technological solutions available. The rest of this article discusses a series of principles scientists should strive to respect when implementing new platform and that have already been implemented with great success in many social web applications that face similar evaluation challenges.

### OPENNESS AND TRANSPARENCY

To combat the opacity of the current peer review system, openness and transparency should be central design features of any next-generation scientific evaluation platform. In this context, openness doesn’t just mean making reviews of papers accessible online; it implies a fundamental level of transparency and data accessibility that should reside at the very core of new platforms. Multiple layers of information—including nearly all the data amassed by that platform over time—should be freely available and programmatically accessible to interested parties.

### OPEN ACCESS TO (NEARLY) ALL CONTENT

Arguably the single most important desideratum for a next-generation evaluation platform is providing open access to the reviews, comments, and ratings of manuscripts generated at all stages of the evaluation process. Setting aside for the moment the question of whether reviewers should be forced to disclose their identities (see below), there is little reason to withhold the content of reviews and ratings from the public—at least in aggregate form (e.g., providing the mean rating of each manuscript). Making evaluations openly accessible would have several substantial benefits. First, it would allow researchers to evaluate the evaluators; that is, researchers would be able to determine the quality of the reviews that influence the reception of an article, and adjust that reception accordingly. Unscrupulous researchers would, for instance, no longer have the power to reject competitors’ work by providing excessively negative reviews, since those reviews would themselves be subject to evaluation. Second, when implemented on a sufficiently large-scale, an open database of evaluations would provide a centralized forum for discussion of scientific work, which currently occurs in a piecemeal and much less efficient fashion elsewhere online and offline. Third, open access to reviews would allow researchers to receive credit for



evaluating others' work, and hence provide greater incentive to participate in peer review.

All three of these principles are already embodied in many existing community-oriented websites. One particularly effective example is implemented on the popular social news website reddit (reddit.com), which features threaded conversations that allow users to comment and vote on both original submissions and other users' comments. Submissions and comments can then be sorted in a variety of ways (e.g., by top score, novelty, by amount of controversy, etc.). The result is a highly efficient collaborative filtering system (Schafer et al., 2007) that rapidly differentiates between high- and low-quality submissions. Moreover, the comments exert a strong influence on the reception of the original submissions; in many cases, an astute comment or two (e.g., when critical questions are raised about the veracity of information provided in a link) leads to rapid adjustment of a submission's score. And since comments are themselves subject to evaluation, the process is iterative and encourages genuine discussion between users with differing opinions. The net result is an openly accessible record of (mostly) intelligent debate over everything from YouTube videos to government bills to old photographs. The same type of open discussion model could potentially greatly facilitate evaluation of scientific manuscripts.

### TRANSPARENT IDENTITIES

While there appear to be few downsides to making the *content* of reviews and ratings openly and easily accessible within a post-publication framework, the question of whether to force disclosure of reviewers' identities is a more delicate one. There's a common perception that peer reviewers would refuse to review papers if forced to disclose their identities, and that anonymous reviews are a necessary evil if we want researchers to express their true views about manuscripts (Fabiato, 1994; Ware, 2007; Baggs et al., 2008). This perception appears to be unfounded inasmuch as empirical studies suggest that forcing reviewers to disclose their identities to authors and/or readers only modestly increases refusal rates while improving the tone of reviews and leaving their overall quality unaltered (Justice et al., 1998; van Rooyen et al., 1998; Walsh et al., 2000; van Rooyen et al., 2010). Moreover, one can legitimately question whether anonymity currently allows reviewers to go to the opposite extreme, expressing excessively negative or unfair views that the light of day might otherwise moderate.

Nonetheless, privacy concerns deserve to be taken seriously. We can distinguish between technical and sociological questions related to identity disclosure. From a technical standpoint, the principle is clear: any evaluation platform should build in tools that allow users a range of privacy management options, ranging from full disclosure of identity (including real names, institutional affiliations, etc.) to pseudonymous or entirely anonymous posting. The sociological question will then arise as to how much transparency of identity is desirable, and how to best motivate that degree of disclosure. A strong case can be made that some data should remain private by default (except in the aggregate); for instance, it would probably be a bad idea to force public display of users' ratings of individual articles. While greater transparency may generally be a good thing, we shouldn't let the

perfect be an enemy of the good: if the only way to encourage widespread adoption of a next-generation evaluation platform is to allow pseudonymity or anonymity that seems preferable to building an idealistic platform that no one wants to use. And as I discuss in more detail below, there is good reason to believe that given a well-structured reputation management system, most users would voluntarily opt to disclose their identities.

### PUBLIC APIs

Application programming interfaces (APIs) play a central role in modern web applications. Public APIs allow third-party developers and users to plot custom bicycle routes on Google Maps, to "mashup" different YouTube videos, and to integrate Twitter streams into their own websites and applications. API-based access to the data generated by a successful scientific evaluation platform would facilitate the development of novel third-party applications, in turn spurring greater adoption of a platform and promoting further innovation. Given a platform that aggregates citation data, ratings, reviews, and comments for every paper in PubMed, and makes such data accessible via API, third party developers could build a broad range of applications—for instance, article recommendation tools ("users who liked this paper also liked these ones..."), specialized aggregators that selectively highlight a subset of articles defined by some common interest, and customizable evaluation metrics that allow users to generate their own weighting schemes for quantitative assessment of articles, journals, researchers, or institutions.

Although the deployment and adoption of research-related APIs is still in early stages, several services have already begun to provide public API access to their data. Notable examples are the Public Library of Science (PLOS) API (<http://api.plos.org>), which provides access to article-level metrics (e.g., page views and downloads) for tens of thousands of PLoS articles, and the Mendeley API (<http://dev.mendeley.com>), which provides programmatic access to a crowdsourced research database of over 100 million articles and growing. An explicit goal of these APIs—and in the case of Mendeley, of an accompanying release of usage data for nearly 5 million papers (<http://dev.mendeley.com/datachallenge>)—is to support development of new research tools such as article recommendation systems (discussed in the next section). These releases represent only the beginning of what promises to be a deluge of publicly accessible data relevant to the evaluation of scientific output.

### PERSONALIZATION AND CUSTOMIZABILITY

There was a time not too long ago when people decided what movies to watch, or what music to listen to, largely on the basis of consensus opinion and/or the authoritative recommendation of a third party. While such factors still play an important role in our choice of media, they have, in many cases, been superseded by social web applications explicitly designed to provide personalized recommendations based on each individual's prior history and preferences. Sophisticated recommendation systems at the heart of many of the web's most popular sites (e.g., Netflix, Amazon, Last.fm, and Google News) now provide nearly effortless ways to identify new products and services we (as opposed to other people) are likely to enjoy (for review, see Adomavicius



and Tuzhilin, 2005; Pazzani and Billsus, 2007; Schafer et al., 2007). The revolutionary impact of such systems lies in their recognition that what people predominantly care about is how much *they* like a product. Other people's evaluations, while informative, are generally helpful only to the extent that they provide a reasonable proxy for one's own preferences.

Broadly speaking, recommendation systems come in two flavors. *Collaborative filtering* approaches rely on user-provided ratings to generate recommendations (Schafer et al., 2007). Make a few 5-point ratings on Netflix, and you'll start receiving suggestions for movies that similar users liked; view a product on Amazon, and it'll try to sell you related products others have bought. *Content-based* approaches rely on objective coding of different aspects of a product or service in order to identify similar items (Pazzani and Billsus, 2007). For example, the Pandora music service bases its recommendations on expert ratings of hundreds of thousands of songs (Casey et al., 2008). Empirical studies demonstrate that both collaborative filtering and content-based recommendation systems—as well as many hybrid approaches—are capable of accurately predicting user preferences across a broad range of domains, including commercial products (Pathak et al., 2010; Sarwar et al., 2000), movies (Miller et al., 2003; Bennett and Lanning, 2007), news articles (Phelan et al., 2009; Liu et al., 2010), leisure activities (Ducheneaut et al., 2009), and musical tastes (Yoshii et al., 2008; Barrington et al., 2009).

In principle, the scientific community could use similar filtering approaches to evaluate scientific output. The fundamental challenge time-pressed researcher's face when evaluating the scientific literature closely resembles the one that consumers in other domains face—namely, how to filter an unmanageable amount of information down to only those items that are likely to be of substantive interest. Currently, scientists address this problem using heuristics of varying quality, e.g., by focusing on highly-cited papers that appear in prestigious journals, signing up for keyword alerts, performing targeted literature searches, and so on. Such approaches can work well, but they're time consuming and effortful. Recommendation systems offer what is, in principle, a superior alternative: instead of requiring explicit effort to identify items of potential interest, the system continuously mines an accumulated database of article metadata and user ratings to generate recommendations. Preliminary efforts using content-based (Dumais and Nielsen, 1992; Basu et al., 2011), collaborative filtering (Bogers and Van Den Bosch, 2008; Naak et al., 2009), or hybrid (Torres et al., 2004; Gipp et al., 2009) approaches demonstrate the viability of automatically generating article recommendations. However, to date, such efforts have been conducted on a small scale, and lack an online, publicly accessible implementation with sufficient appeal to attract a critical mass of users. Developing an integrated recommendation system should thus be a major design goal of next-generation scientific evaluation platforms.

A successfully implemented article recommendation system would reduce researchers' reliance on other heuristics of debatable utility; for instance, given a system that could accurately predict which articles a user would find relevant and of high quality, there would be less need to focus attention on the journals in

which articles were published. The goal of such recommendation systems wouldn't be to serve as final arbiter of the quality of new publications, but simply to filter the literature to a sufficient degree that researchers could efficiently finish the job. Moreover, as discussed in the next section, the presence of a recommendation system would provide a valuable incentive for users to contribute their own evaluations and ratings, enabling an evaluation platform to grow much more rapidly. Naturally, new concerns would arise during the course of implementation; for example, a recommendation system that attempts to identify papers that users will like risks creating an “echo chamber” where researchers only receive recommendations for papers that concord with their existing views (Massa and Avesani, 2007). However, such challenges should generally have straightforward technical solutions. For example, the echo chamber effect could be combated by limiting the weighting of users' favorability ratings relative to other criteria such as relevance of content, methodological rigor (as assessed by the entire userbase), and so on.

A second benefit that highly centralized, open access evaluation platforms would afford is the ability to develop customizable new metrics quantifying aspects of scientific performance that are currently assessed primarily subjectively. Consider, for instance, the task that confronts academic hiring committees charged with selecting a candidate from among dozens or hundreds of potential applicants. Since few if any committee members are likely to have much expertise in any given applicant's exact area of research, hiring decisions are likely to depend on a complex and largely subjective blend of factors. Is an applicant's work well respected by established people in the same field? Does she consistently produce high-quality work, or are many of her contributions incremental and designed to pad her CV? Does a middling citation rate reflect average work, influential work in a small field, or poor work in a large field? Is the applicant's work innovative and risky, or cautious and methodical?

Current metrics don't answer such questions very well. But a centralized and automated evaluation platform could support much more sophisticated quantitative assessment. For instance, a researcher's reputation among his or her peers could be directly quantified using explicit reputation systems (discussed in the next section) based on thousands of data points rather than three self-selected letters of recommendation. The novelty or distinctiveness of a researcher's individual publications could be assessed using algorithms that evaluate similarity of content across articles, pattern of citations to and from other articles, co-authorship, etc., thereby counteracting the pressure many scientists feel to maximize publication rate even if it results in redundant publications (Broad, 1981; Jefferson, 1998; Von Elm et al., 2004). The relative strengths and weaknesses of a research program could be measured by aggregating over users' dimensional ratings of innovation, methodological rigor, clarity, etc. And all of these metrics could be easily normalized to an appropriate reference sample by automatically selecting other authors in the system who works in similar content areas.

Developing an array of such metrics would be an ambitious project, of course, and might be beyond the capacity of any single organization given that funding for such a venture seems

likely to come primarily from the public sector. But the public availability of rich APIs would off-load much of the workload onto motivated third parties. The recent proliferation of metrics such as the h-index (Hirsch, 2005), g-index (Egghe, 2006), m-index (Bornmann et al., 2008), and dozens of other variants (Bornmann et al., 2011) is a clear indicator that a large market exists for better measures of research performance. But such metrics are currently based almost entirely on citation counts; developing a centralized and open platform that supports much richer forms of evaluation (votes, ratings, reviews, etc.) seems likely to spur a broader revolution in bibliometrics (cf. Neylon and Wu, 2009; Lane, 2010; Priem and Hemminger, 2010).

In the longer term, the development of a broad range of evaluation metrics could lead to sophisticated new weighting schemes optimized for highly specific evaluation purposes. Instead of relying solely on recommendation systems to identify relevant articles, researchers would be able to explicitly manipulate the algorithms that generate summary evaluations of both individual articles and researchers' entire output. For instance, a hiring committee could decide to emphasize metrics assessing innovation and creativity over methodological rigor, or vice versa. An editorial board at a general interest journal could use metrics quantifying breadth of interest (e.g., diffusion of positive ratings across researchers from different fields) to select preprints for "official" publication. Science journalists could preferentially weight novelty when selecting work to report on. The degree of customization would be limited only by the sophistication of the underlying algorithms and the breadth of the available research metrics.

Providing a high degree of personalization and customizability wouldn't completely eliminate subjective criteria from evaluation decisions, of course—nor should it. But it would minimize the intensive effort researchers currently invest in filtering the literature and identifying relevant studies; it would reduce reliance on evaluation heuristics of questionable utility (e.g., identifying the quality of papers with the impact factor of journals); and it would provide objective bases for decisions that currently rely largely on subjective criteria. In view of the low reliability of classic peer review, and the pervasive finding that trained human experts are almost invariably outperformed by relatively simple actuarial models (Dawes et al., 1989; White, 2006; Hanson and Morton-Bourgon, 2009), we have every reason to believe that increasing the level of automation and quantitative measurement in the evaluation process will pay large dividends. And there is little to lose, since researchers would always remain free to fall back on conventional metrics such as citation rates if they so desired.

## PROVIDING APPROPRIATE INCENTIVES

Suppose one implemented a platform with features such as those described in the preceding sections. Would scientists rush to use it? Would the database quickly fill up with lengthy reviews and deep comment threads? Probably not. Technical innovation is only one part of any novel publishing platform—and arguably not the most important part. New tools and platforms are often adopted quite slowly, even when they offer significant technical advantages over previous approaches. Users signing up for a service are generally not interested in what the service *could* be like

in five years given widespread adoption; they're interested in the benefits they can obtain from the service if they start using it *right now*.

Many technically advanced platforms that could in principle enhance scientific communication and evaluation fail to appropriately incentivize their potential userbase. Consider the PLoS platform (<http://plos.org>), which has long enabled users to rate and review papers, with the goal of encouraging interaction between readers and/or authors. In theory, such a platform offers substantial benefits to the scientific community. If everyone used it regularly, it would be very easy to tell what other people—including leading experts in the field—thought about any given article. Unfortunately, the PLoS platform provides virtually no incentive to participate, and may even offer disincentives (Neylon and Wu, 2009; Nielsen, 2009). At present, if I spend an hour or two writing a critical review of a paper and sign it with my real name, very few people are likely to read my commentary—and those who do may well wonder why I'm wasting my time writing lengthy reviews on open access websites when I could be working on my own papers. As a consequence, only a small proportion of PLoS articles have received any comments, and a similar lack of engagement characterizes most other publishing platforms that provide a facility for online discussion of manuscripts (Neylon and Wu, 2009; Gotzsche et al., 2010).

Some critics have seized on the lack of community engagement as evidence of the flaws of a post-publication evaluation model (Poynder, 2011). But the reason that researchers haven't flocked to comment on PLoS articles seems very much like the reason editors often complain about how hard it is to find peer reviewers: there simply isn't any meaningful incentive to contribute. Getting researchers to invest their time building an online portfolio isn't only (or even primarily) about providing the *opportunity* to engage in online discussion; it's also about providing appropriate motivation.

As with many of the other problems discussed above, social web applications have already addressed—and arguably solved—the challenge of incentivizing a userbase to participate. Indeed, virtually every website that relies on user-generated product ratings and reviews faces much the same challenge. For instance, Netflix's business model depends partly on its ability to find you movies that you'll enjoy. That ability, in turn, depends on sophisticated quantitative modeling of movie ratings provided by Netflix users. Without the ratings, Netflix wouldn't be able to tell you that you're likely to enjoy *All About My Mother* if you enjoyed *Spirited Away*. But Netflix users don't rate movies out of an abiding respect for Netflix's bottom line; they rate movies so that Netflix can give them personalized movie recommendations. Netflix doesn't have to ask its users to behave charitably; it simply appeals directly to their self-interest. Analogous models are everywhere online: tell Last.fm or Grooveshark which songs you like, and they'll tailor the songs they play to your preferences; buy a product from Amazon, and it'll try to sell you related products others have bought; upvote a link on reddit and you get to exert direct (if weak) social influence on the community. Not only is the long-term goal—whether making money or building an online community—not emphasized on these websites; it's largely invisible.

*A priori*, it seems reasonable to expect the same type of model to work equally well for scientific evaluation. Many scientists decline invitations to review manuscripts because they can't spare a few hours on relatively thankless labor, but few scientists would be too busy to make a single 5-point rating after reading a paper—especially if it doing so helped the system recommend new papers. The long-term goal of creating a centralized platform for evaluation of scientific manuscripts wouldn't require much emphasis; done right, researchers would be happy to use the service simply as a recommendation engine or bibliography management tool. More sophisticated features (e.g., separate ratings along dimensions such as impact, innovation, and methodological rigor; threaded ratings and reviews of other reviews; etc.) could then be added incrementally without disrupting (and indeed, generally increasing) the appeal of the core platform.

Notably, at least one popular service—Mendeley ([mendeley.com](http://mendeley.com))—already appears to be taking precisely this kind of “passive” approach to community building. Initially billed as a web-based bibliography management tool, Mendeley recently introduced a public API that provides access to its data, and has already begun to add social networking features and statistical reports that could soon form the basis for a community driven recommendation system (<http://dev.mendeley.com>). Crucially, Mendeley has been able to grow its enormous crowdsourced database (over 1 million members and 100 million document uploads as of July, 2011) simply by providing an immediately valuable service, without ever having to appeal to its users' altruism. The success of this model demonstrates that the same principles that have worked wonders for commercial services like Netflix and Last.fm can be successfully adapted to the world of scientific evaluation. The challenge lies not so much in getting users to buy into long-term objectives that benefit the scientific community as a whole, but rather, in making sure that the short-term incentives that *do* drive initial user engagement are naturally aligned with those longer-term objectives.

## REPUTATION MANAGEMENT

Providing short-term incentives such as personalized recommendations can help a platform get off the ground, but in the long run, building and maintaining an active community is likely to require additional incentives—ideally, the same ones that already drive scientific contributions offline. One prominent motivator is reputation. Currently, the primary mechanisms for building a reputation in most fields of science are tangible products such as journal publications, research grants, and conference presentations. Many other contributions that play essential roles in driving scientific progress—e.g., peer review, data sharing, and even informal conversation over drinks—historically haven't factored much into scientists' reputations, presumably because they've been difficult to track objectively. For instance, most scientific articles already include extensive discussion and evaluation of prior work—the quality of which bears directly on an author's reputation—but there is currently no way to formally track such embedded discussions and credit authors for particularly strong (or poor) evaluations. The development of new evaluation platforms will make it easy to quantitatively measure, and assign credit for, such contributions. The emerging challenge will be to

ensure that such platforms also provide sufficient incentives for researchers to engage in desirable but historically underappreciated behaviors.

Here, again, scientists can learn from the social web. Reputation systems are at the core of many popular social web communities, including a number that cater explicitly to scientists. A common feature of such communities is that users can endorse or rate other users' contributions—e.g., indicating whether comments are helpful, whether product reviews are informative, and so on. A particularly relevant model is implemented on Stack Exchange (<http://stackexchange.com>), a network of over 50 question and answer sites geared toward professionals in different areas. While the most popular SE website (Stack Overflow) caters to software developers, the network also features a number of popular Q&A sites populated by academic researchers, including mathematics, statistics, physics, and cognitive science exchanges. A key feature of the SE platform is the use of a point-based reputation system. Users receive and award points for questions, answers, and edits that receive favorable ratings from other users. In addition to providing an index of each user's overall contribution to the site, users attain additional privileges as they gain reputation—e.g., the ability to promote, edit, or moderate others' questions. Thus, the system incentivizes users to participate in prosocial activities and penalizes unhelpful or low-quality contributions.

A notable feature of the SE platform is the explicit encouragement for users to post under their real names so as to leverage (and build) their offline reputations. This is most apparent on MathOverflow (<http://mathoverflow.net>), where many prominent users are tenured or tenure-track professors in mathematics-related fields at major research universities—many at the top of their fields. The success of this model demonstrates that, given the right incentives, even busy academics are willing to engage in online activities that, despite their obvious value to the community, previously weren't viewed as creditable scientific contributions. Consider a telling quote from a recent Simons Foundation article (Klarreich, 2011):

“I have felt the lure of the reputation points,” acknowledges Fields medalist Timothy Gowers, of Cambridge University. “It's sort of silly, but nevertheless I do get a nice warm feeling when my reputation goes up.”

Prior to the introduction of collaborative platforms like Stack Exchange, one might have been understandably skeptical of a famous mathematician revealing that he spends much of his time accumulating virtual points online (and as of this writing, Gowers ranks as one of the top 20 users on MathOverflow). But when the points in question are awarded for prosocial activities like asking and answering research questions, reviewing others' work, providing data, writing software, and giving advice, the scientific community stands to reap large benefits. Moreover, in addition to incentivizing prosocial contributions, SE-like reputation systems provide at least two other benefits. First, the reputation scores generated by platforms like Stack Overflow are themselves valuable in evaluating users' contribution to the scientific community, since a high reputation score by definition denotes a user who has made many positive contributions to the scientific community—mostly through channels that established

metrics like citation counts don't adequately assess. Second, the ability to assign credit for contributions outside the traditional scope of scientific publication should incentivize contributions from many people who currently lack the means to contribute to science in more conventional ways. In particular, trained scientists who work at teaching positions or in non-academic settings would have a way of contributing in a meaningful and creditable way to the scientific enterprise even if they lack the time and resources to produce original research. Thus, carefully designed reputation systems stand to have a transformative effect on the communication and evaluation of scientific output.

## WHAT HAPPENS TO TRADITIONAL PRE-PUBLICATION REVIEW?

Supposing new technological platforms do eventually transform the scientific evaluation process, an important outstanding question concerns the role of the traditional, journal-based evaluation model centered on pre-publication review. What happens to this model in a world populated by the kind of evaluation platforms envisioned here? Broadly speaking, there are two potential answers. First, one can envision hybrid evaluation models that combine the best elements of closed/pre-publication review and open/post-publication review. For example, one common argument in favor of pre-publication review is that it improves the quality of a manuscript prior to its public release (Goodman et al., 1994). Although the same benefit could arguably be provided by any platform that allows authors to continually revise their manuscript in response to post-publication reviews, one could certainly opt to retain an element of pre-publication review in an otherwise open platform. A straightforward way to implement such a system would be to grant authors permission over who can view a manuscript. In an initial "closed" period, authors would be free to invite selected peers to perform a closed review of the manuscript. The feedback received could then be used to revise the manuscript until the authors were satisfied. The key point is that the control over when to publish the "official" version of the manuscript would rest with the authors and not with an editor (though one might perhaps force authors to stipulate ahead of time whether or not each review would be made public, ensuring that authors could not suppress negative reviews *post hoc*). A major benefit of such an approach is that it would allow diligent authors to solicit feedback from competing (and likely critical) researchers, while penalizing less careful authors who rush to publish without soliciting feedback first. In contrast, under the current system, recommending critical reviewers is a risky and generally detrimental proposition.

The second way to answer the "what happens to conventional review" question is to admit that we don't really know—and, more importantly, that we don't really have to know. If conventional journals and pre-publication review play an indispensable role in the evaluation process, nothing much should change. Journals could go on serving exactly the same role they presently serve. All of the benefits of next-generation platforms discussed would apply strictly to post-publication review, after the standard review process has run its course (e.g., a deeply flawed article that happened to get by the peer review process at a top-tier journal would be susceptible to immediate and centralized

post-publication critique). There would be no need to expend effort actively trying to eliminate conventional journals; a well-designed evaluation platform should be agnostic with respect to the venue (if any) in which manuscripts originally appear. Moreover, from a pragmatic standpoint, adoption of new post-publication evaluation platforms is likely to occur more rapidly if such platforms are presented as complements to conventional review rather than as competitors.

That said, it's easy to see how sophisticated post-publication evaluation platforms might ultimately obviate any need for conventional journals, and many commentators have argued that this is a perfectly logical and desirable end result (LaPorte et al., 1995; Odlyzko, 1995; Delamothe and Smith, 1999; Kingsley, 2007; Smith, 2010). Once it becomes clear that one can achieve efficient and reliable evaluation of one's manuscripts regardless of where (or whether) they're officially published, there will be little incentive for authors to pursue a traditional publication route. As a result, traditional journals may simply disappear over time. But the important point is that if this process happens, it will happen organically; nothing about the type of platform proposed here explicitly constrains the role of journals in any way. To the extent that traditional journals offer scientists an irreplaceable service, they will presumably continue to thrive. And if they don't offer a valuable service, we shouldn't mourn their passing.

## PUTTING IT INTO PRACTICE

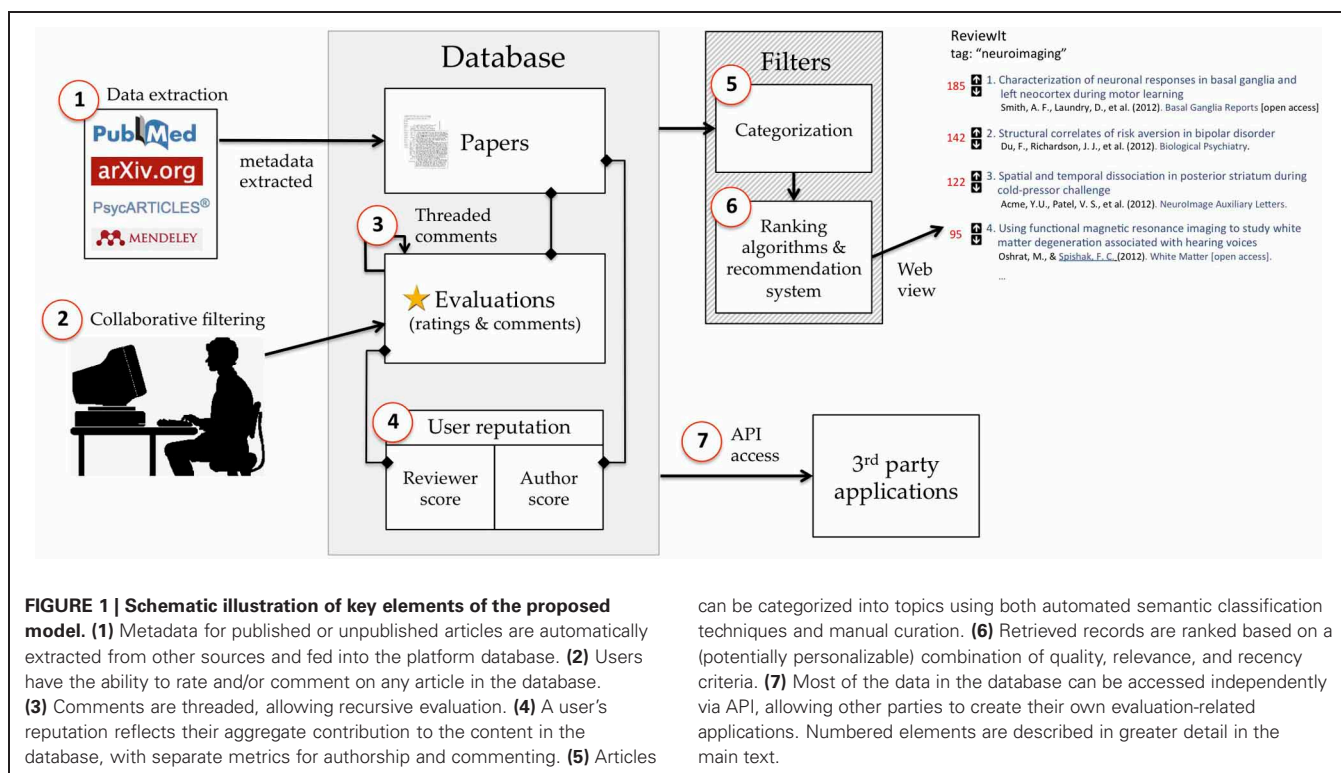
Having reviewed a number of basic design considerations, this section outlines one possible specification for a post-publication evaluation platform. In contrast to a number of recent proposals that focus on wholesale restructuring of the scientific publishing and evaluation process (e.g., Pöschl, 2010; Kravitz and Baker, 2011; Nosek and Bar-Anan, 2012), the platform described here focuses exclusively on facilitating centralized, publisher-independent post-publication review. The platform would exist independently of the existing pre-publication review system, and would not require articles to have undergone any prior form of peer review before being added to the system. Thus, there are effectively no major technical or legal barriers (e.g., copyright restrictions) to the immediate implementation of such a platform, and social barriers are also minimized by presenting the platform as a complement rather than competitor to traditional models.

A schematic of the proposed platform is provided in **Figure 1**. Given that the central argument of this paper is that most of the principles needed to establish a successful evaluation platform are already widely implemented on the social web, it should come as no surprise that the platform described here features few novel features—it's essentially a Reddit clone, with a few additional features borrowed from other platforms like Stack Exchange, Netflix, and Amazon. The platform features the following elements (corresponding to the circled numbers in **Figure 1**).

### DATA EXTRACTION

The database is initially populated (and continuously updated) by pulling data from academic search engines and repositories. For example, many services like PubMed and ArXiv.org provide API access or free data dumps, ensuring that the evaluation platform can remain up to date without requiring any user input.





The evaluation platform should link to all articles on the original publisher/repository website (when available), but should not take on the responsibility of facilitating access to articles. Articles that are currently behind a pay-wall would not become publicly accessible in virtue of having a discussion page on the evaluation platform website; the system would (at least initially) operate in parallel with the traditional publishing system rather than in competition.

### COLLABORATIVE FILTERING

At the core of the platform is a collaborative filtering approach that allows any registered user to rate or comment on any article in the database. These ratings and comments can then be used to sort and rank articles and users in a variety of ways (see below). The simplest implementation would be a reddit-like voting system that allows users to upvote or downvote any article in the database with a single click (Figure 2A). More sophisticated approaches could include graded ratings—e.g., 5-point responses, like those used by Amazon or Netflix—or separate rating dimensions such as methodological rigor, creativity and innovation, substantive impact, etc., providing users with an immediate snapshot of the strengths and weaknesses of each article.

### THREADED COMMENTING

A key feature of the reddit platform (and many of its precursors, e.g., Slashdot.org) is threaded discussion: users can comment on and rate not only on primary documents (in this case, scientific articles), but also other comments (Figure 2B). This feature is vital to the success of a collaborative filtering platform, as it provides a highly efficient corrective mechanism.

For example, it is common on reddit to see one comment's score change dramatically in a span of hours in response to additional comments. This format should translate exceptionally well to the domain of scientific evaluation, where a single user has the potential to raise important concerns that other researchers may have overlooked but can nonetheless appreciate. To encourage authors to engage with other commenters, one might designate verified author comments with a special icon (e.g., Figure 2B, orange), and perhaps provide a small ratings boost to such comments.

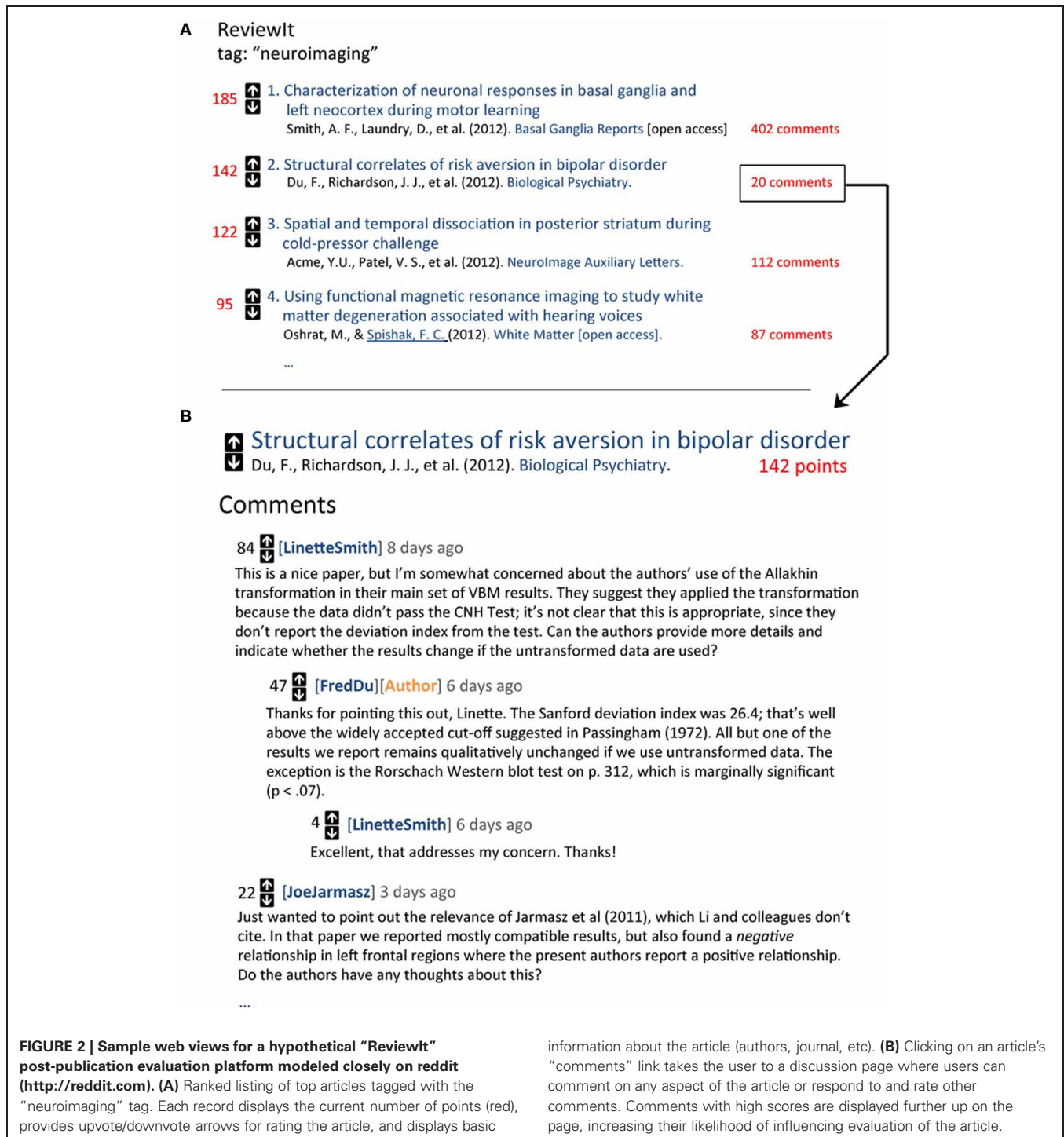
### REPUTATION SYSTEM

To incentivize users to comment on and review papers (and participate in threaded discussions of those reviews), the evaluation platform should feature a robust reputation system that combines basic features of the reddit system with additional features found in the Stack Exchange platform. The reddit system awards users "karma" points for sharing links and writing comments that are favorably rated by other users; in the context of a scientific evaluation platform, users would receive points based on ratings of their contributed articles on the one hand and their comments and reviews on the other (these commenting and authorship metrics would be kept separate). Each user's reputation and a summary of their contributions would be viewable from the user's public page. Standard filtering and search options would be available, allowing other users to see, e.g., what an individual's top-rated or newest articles or comments are.

### CATEGORIZATION

Much as reddit features "subreddits" geared toward specific topics and niche interests (e.g., science, cooking, or politics), articles





in the database would be organized by topic. A core set of topics could be automatically generated based on keywords (e.g., the National Library of Medicine's Medical Subject Heading [MeSH] ontology); thus, for example, navigating to the /keyword/neuroimaging subdirectory would display a ranked list of all articles tagged with the "neuroimaging" keyword (e.g., **Figure 2**). Additionally, however, users would be able to create their own custom topics tailored to more specific niches, much as any reddit

user currently has the ability to create new subreddits. This two-pronged approach would balance the need for relatively objective ontologies with manually curated sets of articles (where the role of curator would be somewhat similar to that of an editor in the conventional publishing system). An additional benefit of topic-based organization is that articles published in domains with very different citation rates and community sizes could be easily normalized and put on a common metric, much as the reddit front

page currently normalizes scores of links submitted to different subreddits.

## RANKING

For any given set of articles retrieved from the database, a ranking algorithm would be used to dynamically order articles on the basis of a combination of quality (an article's aggregate rating in the system), relevance (using a recommendation system akin to Netflix or Amazon's), and recency (newly added articles would receive a boost). By default, the same algorithm would be used for all users (as on reddit). However, as discussed above, allowing users to customize the algorithm used to rank articles and/or weight researchers contributions would greatly increase the utility of the basic platform by enabling individuals or groups with specific goals to filter articles or users more efficiently (e.g., faculty search committees with specific needs could rank candidates based on a customized set of criteria).

## API ACCESS

To facilitate community engagement and allow third parties to use evaluation data in creative new ways, a public API should be provided that enables programmatic access to nearly all platform data (with the exception of data where privacy is a potential issue—e.g., individual users' ratings of individual articles).

Importantly, these features need not all be implemented at once. In particular, recommendation systems, customizable ranking algorithms, and a public API, while all desirable, could be added at later stages of implementation once the basic platform was operational. Of course, many other features not mentioned here could also be added later—e.g., social networking features, integration with third-party evaluation metrics (e.g., total-impact.org), a closed-review phase that allows users to solicit reviews privately before an article's public release, and so on.

## CONCLUSION

In the preface to *On Spirals*, Archimedes amusingly reveals that, on at least one occasion, he deliberately sent his colleagues in Alexandria false theorems, “so that those who claim to discover everything, but produce no proofs of the same, may be confuted as having pretended to discover the impossible” (Bombieri, 2011). This age-old concern with being scooped by other researchers will no doubt be familiar to many contemporary scientists. What's not so easily understandable is why, in an age of preprint servers, recommendation systems, and collaborative filters, we continue to employ publication and evaluation models that allow such concerns to arise so frequently

in the first place. While healthy competition between groups may be conducive to scientific progress, delays in the review and publication process are almost certainly not. Inefficiencies in our current evaluation practices are visible at every stage of the process: in the redundancy of writing and re-writing articles in different formats to meet different journals' guidelines; in the difficulty editors face in locating appropriate reviewers; in the opacity and unreliability of the pre-publication review process; in the delays imposed by slow reviews and fixed publication schedules; in limitations on access to published articles; and in the lack of centralized repositories for post-publication evaluation of existing work. Almost without exception, effective technical solutions to these inefficiencies already exist, and are in widespread use on the social web. And yet, almost without exception, the scientific community has ignored such solutions in favor of an antiquated evaluation model that dates back hundreds of years—and in some respects, all the way back to the ancient Greeks.

To take a long view, one might argue that such inefficiencies are not the end of the world; after all, science is a cumulative, self-correcting enterprise (Peirce, 1932; Platt, 1964; Popper, 2002). Given sufficient time, false positives work themselves out of the literature, bad theories are replaced by better ones, and new methods emerge that turn yesterday's tour-de-force analysis into today's routine lab assay. But while the basic truth of this observation isn't in question, it's also clear that all cumulative efforts are not equal; the rate at which we collectively arrive at new scientific discoveries counts for something too. Ideally, we'd like to find cures for diseases, slow the aging process, and build colonies on extra-solar planets sooner rather than later. Since the rate of scientific discovery is closely tied to the rate of dissemination and evaluation of scientific output, the research community has an enormous incentive—and arguably, a moral duty—to improve the efficiency and reliability of the scientific evaluation process. From a utilitarian standpoint, it seems almost certain that even relatively small increases in the rate of scientific publication and evaluation would, compounded over time, have far greater societal benefits than all but a very few original scientific discoveries. We should act accordingly, and not let inertia, lack of imagination, or fear of change prevent us from realizing new models of scientific evaluation that are eminently feasible given present-day technologies.

## ACKNOWLEDGMENTS

The author thanks John Clithero and Brian Cody for helpful comments on an earlier draft of this manuscript.

## REFERENCES

- Adomavicius, G., and Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* 17, 734–749.
- Baggs, J. G., Broome, M. E., Dougherty, M. C., Freda, M. C., and Kearney, M. H. (2008). Blinding in peer review: the preferences of reviewers for nursing journals. *J. Adv. Nurs.* 64, 131–138.
- Barrington, L., Oda, R., and Lanckriet, G. (2009). “Smarter than genius? human evaluation of music recommender systems,” in *Proceedings of the International Society for Music Information Retrieval Conference*, 357–362.
- Basu, C., Cohen, W. W., Hirsh, H., and Nevill-Manning, C. (2011). Technical paper recommendation: a study in combining multiple information sources. *J. Artif. Intell. Res.* 14, 231–252.
- Bath, F. J., Owen, V. E., and Bath, P. M. (1998). Quality of full and final publications reporting acute stroke trials: a systematic review. *Stroke* 29, 2203–2210.
- Bennett, J., and Lanning, S. (2007). “The netflix prize,” in *Proceedings of KDD Cup and Workshop*, Vol. 2007 (San Jose, CA), 8.
- Bogers, T., and Van Den Bosch, A. (2008). “Recommending scientific articles using citeulike,” in *Proceedings of the 2008 ACM Conference on Recommender Systems RecSys 08* (Lausanne, Switzerland), 23, 287.
- Bombieri, E. (2011). “The mathematical infinity,” in *Infinity: New Research Frontiers*, eds M. Heller and W. H. Woodin (Cambridge, UK: Cambridge University Press), 55–75.

- Bornmann, L., Mutz, R., and Daniel, H.-D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *J. Am. Soc. Inf. Sci. Technol.* 59, 830–837.
- Bornmann, L., Mutz, R., and Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: a multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE* 5:e14331. doi: 10.1371/journal.pone.0014331
- Bornmann, L., Mutz, R., Hug, S. E., and Daniel, H.-D. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *J. Informetr.* 5, 346–359.
- Broad, W. (1981). The publishing game: getting more for less. *Science* 211, 1137–1139.
- Casati, F., Giunchiglia, F., and Marchese, M. (2007). Publish and perish: why the current publication and review model is killing research and wasting your money. *Ubiquity* 2007, 3.
- Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). “Content-based music information retrieval: current directions and future challenges,” in *Proceedings of the IEEE* 96, 668–696.
- Chew, F. (1991). Fate of manuscripts rejected for publication in the *AJR*. *Am. J. Roentgenol.* 156, 627–632.
- Chondros, T. G. (2010a). “Archimedes influence in science and engineering,” in *The Genius of Archimedes—23 Centuries of Influence on Mathematics, Science and Engineering*, Vol. 11, eds S. A. Paipetis and M. Ceccarelli (Dordrecht: Springer Netherlands), 411–425.
- Chondros, T. G. (2010b). Archimedes life works and machines. *Mech. Mach. Theory* 45, 1766–1775.
- Dawes, R., Faust, D., and Meehl, P. (1989). Clinical versus actuarial judgment. *Science* 243, 1668–1674.
- Delamothe, T., and Smith, R. (1999). Moving beyond journals: the future arrives with a crash. *BMJ* 318, 1637–1639.
- Dong, P., Loh, M., and Mondry, A. (2005). The “impact factor” revisited. *Biomed. Digit. Libr.* 2, 7.
- Ducheneaut, N., Partridge, K., Huang, Q., Price, B., Roberts, M., Chi, E., Bellotti, V., and Begole, B. (2009). “Collaborative filtering is not enough? Experiments with a mixed-model recommender for leisure activities,” in *User Modeling Adaptation and Personalization*, Vol. 5535, eds G.-J. Houben, G. McCalla, F. Pianesi, and M. Zancanaro (Trento, Italy), 295–306.
- Dumais, S. T., and Nielsen, J. (1992). “Automating the assignment of submitted manuscripts to reviewers,” in *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR 92*, eds N. Belkin, P. Ingwersen, and A. M. Pejtersen (Copenhagen, Denmark: ACM Press), 233–244.
- Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics* 69, 131–152.
- Ellison, G. (2002). The slowdown of the economics publishing process. *J. Polit. Econ.* 110, 947–993.
- Fabiato, A. (1994). Anonymity of reviewers. *Cardiovasc. Res.* 28, 1134–1139.
- Gipp, B., Beel, J., and Hentschel, C. (2009). “Scienstein: a research paper recommender system,” in *Proceedings of the International Conference on Emerging Trends in Computing ICETiC’09*, Vol. 301, (Virudhunagar, Tamilnadu, India: Citeseer), 309–315.
- Goodman, S. N., Berlin, J., Fletcher, S. W., and Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at *Annals of Internal Medicine*. *Ann. Intern. Med.* 121, 11–21.
- Gotzsche, P. C., Delamothe, T., Godlee, F., and Lundh, A. (2010). Adequacy of authors’ replies to criticism raised in electronic letters to the editor: cohort study. *BMJ* 341, c3926.
- Hall, S. A., and Wilcox, A. J. (2007). The fate of epidemiologic manuscripts: a study of papers submitted to epidemiology. *Epidemiology* 18, 262–265.
- Hanson, R. K., and Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: a meta-analysis of 118 prediction studies. *Psychol. Assess.* 21, 1–21.
- Harnad, S. (1999, July 18). Free at last: the future of peer-reviewed journals. *D-Lib Magazine*.
- Heath, T. L. (1897). *The Works of Archimedes*. Cambridge, UK: Cambridge University Press.
- Hirsch, J. E. (2005). An index to quantify an individual’s scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* 102, 16569–16572.
- Hojat, M., Gonnella, J., and Caelleigh, A. (2003). Impartial judgment by the “Gatekeepers” of science: fallibility and accountability in the peer review process. *Adv. Health Sci. Educ.* 8, 75–96.
- Ioannidis, J. P. A., Tatsioni, A., and Karassa, F. B. (2010). Who is afraid of reviewers’ comments? Or, why anything can be published and anything can be cited. *Eur. J. Clin. Invest.* 40, 285–287.
- Jefferson, T. (1998). Redundant publication in biomedical sciences: scientific misconduct or necessity? *Sci. Eng. Ethics* 4, 135–140.
- Jefferson, T., Rudin, M., Brodney Folse, S., and Davidoff, F. (2007). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database Syst. Rev.* 2, MR000016.
- Justice, A. C., Cho, M. K., Winker, M. A., Berlin, J. A., and Rennie, D. (1998). Does masking author identity improve peer review quality? A randomized controlled trial. *PEER Investigators. JAMA* 280, 240–242.
- Kingsley, D. (2007). The journal is dead, long live the journal. *Horizon* 15, 211–221.
- Klarreich, E. (2011). The Global Math Commons. *Simons Foundation*. Retrieved July 17 2011, from [https://simonsfoundation.org/mathematics-physical-sciences/news/-/asset\\_publisher/bo1E/content/the-global-math-commons](https://simonsfoundation.org/mathematics-physical-sciences/news/-/asset_publisher/bo1E/content/the-global-math-commons)
- Kravitz, D. J., and Baker, C. I. (2011). Toward a new model of scientific publishing: discussion and a proposal. *Front. Comput. Neurosci.* 5:1–12. doi: 10.3389/fncom.2011.00055
- LaPorte, R. E., Marler, E., Akazawa, S., Sauer, F., Gamboa, C., Shenton, C., Glosser, C., Villaseñor, A., and Maclure, M. (1995). The death of biomedical journals. *BMJ* 310, 1387–1390.
- Lane, J. (2010). Let’s make science metrics more scientific. *Nature* 464, 488–489.
- Liu, J., Dolan, P., and Pedersen, E. R. (2010). “Personalized news recommendation based on click behavior,” in *Proceedings of the 15th International Conference on Intelligent User Interfaces*. Search ACM (Hong Kong, China), 31–40.
- Mahoney, M. J. (1977). Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cognit. Ther. Res.* 1, 161–175.
- Maier, G. (2006). Impact factors and peer judgment: the case of regional science journals. *Scientometrics* 69, 651–667.
- Massa, P., and Avesani, P. (2007). Trust metrics on controversial users: balancing between tyranny of the majority and echo chambers. *Int. J. Semantic Web Infor. Syst.* 3, 39–64.
- Mayor, J. (2010). Frontiers: are scientists nearsighted gamblers? the misleading nature of impact factors. *Front. Psychol.* 1:215. doi: 10.3389/fpsyg.2010.00215
- Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., and Riedl, J. (2003). “MovieLens unplugged: experiences with an occasionally connected recommender system,” in *Proceedings of the 8th International Conference on Intelligent User Interfaces*, Work ACM (Miami, FL), 263–266.
- Naak, A., Hage, H., and Aïmeur, E. (2009). “A multi-criteria collaborative filtering approach for research paper recommendation in payres,” in *4th International Conference, MCETECH 2009. E-Technologies Innovation in an Open World* (Ottawa, Canada), 25–39.
- Neylon, C., and Wu, S. (2009). Article-level metrics and the evolution of scientific impact. *PLoS Biol.* 7:e1000242. doi: 10.1371/journal.pbio.1000242
- Nielsen, M. (May 2009). Doing science in the open. *Phys. World* 30–35.
- Nosek, B. A., and Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychol. Inq.* 23, 217–243.
- Odlyzko, A. M. (1995). Tragic loss or good riddance? The impending demise of traditional scholarly journals. *Int. J. Hum. Comput. Stud.* 42, 71–122.
- Pathak, B., Garfinkel, R., Gopal, R. D., Venkatesan, R., and Yin, F. (2010). Empirical analysis of the impact of recommender systems on sales. *J. Manag. Inf. Syst.* 27, 159–188.
- Pazzani, M. J., and Billsus, D. (2007). “Content-based recommendation systems,” in *The Adaptive Web*, Vol. 4321, eds P. Brusilovsky, A. Kobsa, and W. Nejdl (Berlin, Germany: Springer), 325–341.
- Pearce, C. S. (1932). *Collected Papers of Charles Sanders Peirce*. Boston, MA: Harvard University Press.
- Phelan, O., McCarthy, K., and Smyth, B. (2009). “Using twitter to recommend real-time topical news,” in *Proceedings of the Third ACM Conference on Recommender Systems RecSys 09*, (New York, NY: Systems’09), 385.
- Platt, J. R. (1964). Strong inference. *Science* 146, 347–353.
- Popper, K. (2002). *The Logic of Scientific Discovery*. Vol. 2, eds M. Archer, R. Bhaskar, A. Collier, T. Lawson, and A. Norrie (New York, NY: The Routledge), 513.
- Poynder, R. (2011). PLoS ONE, open access, and the future of scholarly publishing. Available online at:

- [http://richardpoynder.co.uk/PLoS\\_ONE.pdf](http://richardpoynder.co.uk/PLoS_ONE.pdf)
- Priem, J., and Hemminger, B. H. (2010). Scientometrics 2.0, New metrics of scholarly impact on the social Web. *First Monday*, 15.
- Pöschl, U. (2010). Interactive open access publishing and peer review: the effectiveness and perspectives of transparency and self-regulation in scientific communication and evaluation. *Atmos. Chem. Phys.* 19, 293–314.
- Ray, J. (2000). The fate of manuscripts rejected by a general medical journal. *Am. J. Med.* 109, 131–135.
- Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. *Organization* 5, 158–167.
- Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). Collaborative filtering recommender systems. *Int. J. Electronic Bus.* 2, 77.
- Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ* 314, 497.
- Shadbolt, N., Brody, T., Carr, L., and Harnad, S. (2006). “The open research web: a preview of the optimal and the inevitable,” in *Open Access: Key Strategic, Technical and Economic Aspects*, ed N. Jacobs (Oxford, UK: Chandos Publishing).
- Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *J. R. Soc. Med.* 99, 178–182.
- Smith, R. (2010). Classical peer review: an empty gun. *Breast Cancer Res.* 12(Suppl. 4), S13.
- Sutherland, W. J., Goulson, D., Potts, S. G., and Dicks, L. V. (2011). Quantifying the impact and relevance of scientific research. *PLoS ONE* 6:e27537. doi: 10.1371/journal.pone.0027537
- Torres, R., McNee, S. M., Abel, M., Konstan, J. A., and Riedl, J. (2004). “Enhancing digital libraries with TechLens,” in *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries JCDL 04* (Tucson, Arizona: ACM Press), 228–236.
- Von Elm, E., Poglia, G., Walder, B., and Tramèr, M. R. (2004). Different patterns of duplicate publication: an analysis of articles used in systematic reviews. *JAMA* 291, 974–980.
- Walsh, E., Rooney, M., Appleby, L., and Wilkinson, G. (2000). Open peer review: a randomised controlled trial. *Br. J. Psychiatry* 176, 47–51.
- Ware, M. (2007). *Peer Review in Scholarly Journals: Perspectives of the Scholarly Community – An International Study*. Bristol, UK: Publishing Research Consortium, [Internet].
- West, R., and McIlwaine, A. (2002). What do citation counts count for in the field of addiction? An empirical evaluation of citation counts and their link with peer ratings of quality. *Addiction* 97, 501–504.
- White, M. J. (2006). The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction stefania aegisdottir. *Couns. Psychol.* 34, 341–382.
- Yoshii, K., Goto, M., Komatani, K., Ogata, T., and Okuno, H. G. (2008). An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. *IEEE Trans. Audio Speech Lang. Process.* 16, 435–447.
- Young, N. S., Ioannidis, J. P. A., and Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Med.* 5:e201. doi: 10.1371/journal.pmed.0050201
- van Rooyen, S., Godlee, F., Evans, S., Smith, R., and Black, N. (1998). Effect of blinding and unmasking on the quality of peer review: a randomized trial. *JAMA* 280, 234–237.
- van Rooyen, S., Delamothe, T., and Evans, S. J. (2010). Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial. *BMJ* 341, c5729.

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 October 2011; accepted: 03 September 2012; published online: 01 October 2012.

Citation: Yarkoni T (2012) Designing next-generation platforms for evaluating scientific output: what scientists can learn from the social web. *Front. Comput. Neurosci.* 6:72. doi: 10.3389/fncom.2012.00072

Copyright © 2012 Yarkoni. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.





# Open evaluation: a vision for entirely transparent post-publication peer review and rating for science

Nikolaus Kriegeskorte\*

Medical Research Council, Cognition and Brain Sciences Unit, Cambridge, UK

**Edited by:**

Diana Deca, Technische Universität München, Germany

**Reviewed by:**

Satrajit S. Ghosh, Massachusetts Institute of Technology, USA  
Razvan V. Florian, Romanian Institute of Science and Technology, Romania

**\*Correspondence:**

Nikolaus Kriegeskorte, Medical Research Council, Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, UK.  
e-mail: nikolaus.kriegeskorte@mrc-cbu.cam.ac.uk;  
www.mrc-cbu.cam.ac.uk/people/nikolaus.kriegeskorte/

The two major functions of a scientific publishing system are to provide *access* to and *evaluation* of scientific papers. While open access (OA) is becoming a reality, open evaluation (OE), the other side of the coin, has received less attention. Evaluation steers the attention of the scientific community and thus the very course of science. It also influences the use of scientific findings in public policy. The current system of scientific publishing provides only journal prestige as an indication of the quality of new papers and relies on a non-transparent and noisy pre-publication peer-review process, which delays publication by many months on average. Here I propose an OE system, in which papers are evaluated post-publication in an ongoing fashion by means of open peer review and rating. Through signed ratings and reviews, scientists steer the attention of their field and build their reputation. Reviewers are motivated to be objective, because low-quality or self-serving signed evaluations will negatively impact their reputation. A core feature of this proposal is a division of powers between the accumulation of evaluative evidence and the analysis of this evidence by paper evaluation functions (PEFs). PEFs can be freely defined by individuals or groups (e.g., scientific societies) and provide a plurality of perspectives on the scientific literature. Simple PEFs will use averages of ratings, weighting reviewers (e.g., by *H-index*), and rating scales (e.g., by relevance to a decision process) in different ways. Complex PEFs will use advanced statistical techniques to infer the quality of a paper. Papers with initially promising ratings will be more deeply evaluated. The continual refinement of PEFs in response to attempts by individuals to influence evaluations in their own favor will make the system ungameable. OA and OE together have the power to revolutionize scientific publishing and usher in a new culture of transparency, constructive criticism, and collaboration.

**Keywords:** peer review, publishing, ratings, social web, open evaluation

## INTRODUCTION

A scientific publication system needs to provide two basic functions: access and evaluation. Access means we can read anything, evaluation means we do not have to read everything. The traditional publication system restricts the access to papers by requiring payment, and it restricts the evaluation of papers by relying on just 2–4 pre-publication peer reviews and by keeping the reviews secret. As a result, the current system suffers from a lack of quality and transparency of the peer-review evaluation process, and the only immediately available indication of a new paper's quality is the prestige of the journal it appeared in.

Open access (OA) is now widely accepted as desirable and is in the process of becoming a reality (Harnad, 2010). However, the other essential element, evaluation, has received less attention. The current peer-review system has attracted much criticism (Smith, 2006, 2009; Ware, 2011). Arguments (Smith, 1999; Godlee, 2002; Frishauf, 2009; Boldt, 2010) and experiments (Harnad, 1997; Walsh et al., 2000; Greaves et al., 2006; Pulverer, 2010; Pöschl, 2010) with open review and post-publication commentary have suggested that a more transparent system might have potential. However, we have yet to develop a coherent shared vision for “open evaluation” (OE), and an OE movement comparable to the OA movement.

The evaluation system steers the attention of the scientific community and, thus, the very course of science. For better or worse, the most visible papers determine the direction of each field and guide funding and public policy decisions. Evaluation, therefore, is at the heart of the entire endeavor of science. As the number of scientific publications explodes, evaluation and selection will only gain importance. A grand challenge of our time, therefore, is to design the future system, by which we evaluate papers and decide which ones deserve broad attention. OE, an ongoing post-publication process of transparent peer evaluation (including written reviews and ratings of papers), promises to address the problems of the current system.

Here I outline a vision for an open publication and evaluation system with the following key features: Papers are evaluated in an ongoing fashion after publication by means of reviews and ratings. Reviews are mini-publications and can be signed or anonymous. Signed reviews and signed ratings both contribute to a scientist's visibility. More important papers are more deeply evaluated as they will receive more evaluations. Scientists are more motivated to perform reviews, because it helps build their reputation. Multiple paper evaluation functions (PEFs), freely defined by individuals or groups (e.g., scientific societies, private, and public organizations) provide a plurality of perspectives on the scientific literature. The



transition toward a future system of instant publication can be achieved by providing an OE system that will initially serve to more deeply evaluate important papers published under the current system of pre-publication peer review. When the OE system has proven its superiority to the current system of peer review, it will replace the current system.

First, I briefly describe key features of the current system of scientific publishing and where it falls short. Second, I briefly describe some positive current developments that represent steps in the right direction, but do not go far enough. Third, I present a general vision for scientific publishing, based on OA and OE, using entirely transparent post-publication reviews and ratings and freely definable PEFs. Fourth, I describe a specific plan for a minimalist OE system that is simple and yet could go a long way toward providing the key functionality for accumulating the evaluative evidence. Fifth, I describe a specific plan for a PEF, so as to illustrate more concretely how the accumulated evidence can be combined to prioritize the literature. Sixth, I outline the ultimate goal, free instant scientific publishing with OA and OE. Finally, in the discussion, I address a number of concerns and counter-arguments that have frequently come up in informal discussions. These concerns include a lack of evaluations and the question of how we might smoothly transition toward the envisioned system.

### THE CURRENT SYSTEM OF SCIENTIFIC PUBLISHING

The current system of scientific publishing provides access and evaluation in a limited fashion. While access often requires payment, papers are made available in an appealing professional layout that makes them easier to read. This function is desirable, but not critical to scientific progress. The current system also provides evaluation: It administers peer review and provides an evaluative signal that helps readers choose papers, namely journal prestige. This function is critical to scientific progress. However, journal prestige is a crude measure that is not specific to particular

papers. The overall process of the current system is summarized in **Figure 1**. We will now discuss the main drawbacks.

### THE SYSTEM IS NOT GENERALLY OPEN ACCESS

Scientific papers benefit society only to the extent that they are accessible. If the public pays for scientific research it should demand that the results be openly accessible. If private publishers offer valuable services at reasonable prices that contribute to the dissemination of scientific papers, such as appealing layout, then research institutes may want to purchase them. However, access to results of publicly funded research should never come at a cost to an individual. Since OA is already widely seen as desirable among scientists and the general public, this paper focuses on OE: how to open up the other major function of a publication system, namely the evaluation of scientific papers.

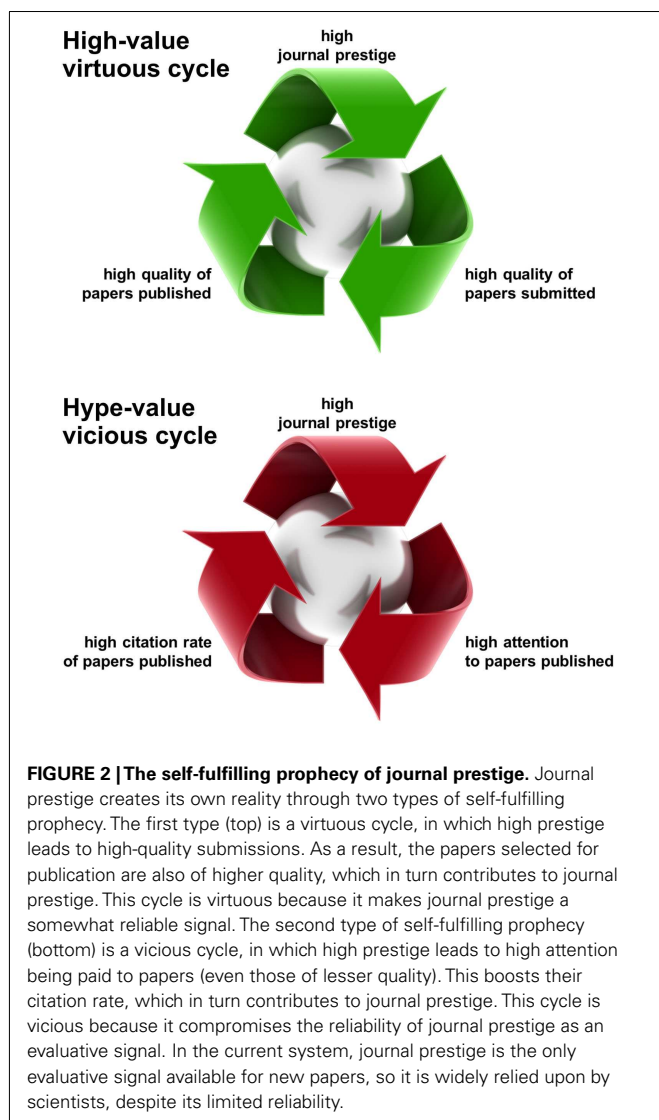
### JOURNAL PRESTIGE, THE ONLY QUALITY INDICATOR FOR NEW PAPERS, PROVIDES AN IMPOVERISHED AND UNRELIABLE EVALUATIVE SIGNAL

The main evaluative signal provided to readers for prioritizing their reading of scientific papers is journal prestige. We are more likely to attend to a paper published in *Nature* than to a similar paper published in a specialized journal. While journal prestige is somewhat correlated with the quality of scientific papers, it is not a reliable indicator of the quality of a particular paper. Moreover, journal prestige as an evaluative signal is compromised by causal circularity: Prestige – once acquired – creates its own reality.

The self-fulfilling prophecy of journal prestige has two component cycles of causality, a virtuous one and a vicious one (**Figure 2**). In the virtuous cycle, prestige brings higher-quality submissions, which in turn contribute to prestige. This cycle is virtuous, because the increase in prestige actually reflects an increase in the quality of the papers. In the vicious cycle, prestige brings higher-citation frequencies (even for average-quality papers), which in turn brings prestige. This cycle is vicious, because it causes journal impact factors (IF) to give a distorted picture of the quality of the published



**FIGURE 1 | The current system.** This flowchart summarizes the process by which the current system operates. Key features include long publication delays, secret peer review, failure to make evaluations (reviews and ratings) available to the community, and journal prestige as the only evaluative signal available immediately upon publication.



papers. IFs and the higher or lesser prestige they confer on journals therefore compromise the public perception of the quality of particular papers.

In addition to being an unreliable indicator of a scientific paper's quality, journal prestige provides only a greatly impoverished, evaluative signal. The detailed reviews and multi-dimensional ratings provided to the journal by the reviewers are kept secret. The reviewers are established experts, largely funded by the public, who work hard to evaluate scientific papers. And yet the detailed evaluations are kept secret and contribute to the reception of a paper only after being reduced to a categorical quality stamp: the journal label. This constitutes a loss to the scientific community and to the general public of valuable judgments that are already being performed and paid for.

#### THE REVIEW PROCESS IS NON-TRANSPARENT, TIME-LIMITED, AND BASED ON TOO FEW OPINIONS

The current system of publishing is based on a non-transparent evaluation process that includes secret reviews visible only to editors and authors. For high-impact publications, the editorial

decision process preceding full review often also includes informal comments solicited by the editors from other scientists. Such informal additional sources of evaluation may often improve the quality of the decisions made – this is why they are used. Nevertheless, this practice compromises the transparency and objectivity of the system.

The selection of a paper for publication is typically based on 2–4 peer reviews. The quality of an original and challenging scientific paper cannot reliably be assessed by such a small number of reviewers – even if the reviewers are experts and have no conflict of interest (i.e., they are not competitors). In reality, the reviewers who are experts in the particular topic of a paper often have some personal stake in the paper's publication. They may be invested in the theory supported or in another theory. More generally, they may have competitive feelings that compromise their objectivity.

For high-impact publications, this political dynamic is exacerbated because the stakes are higher and more scientists are competing for a smaller stage. To make matters worse, high-impact publications require their reviewers to judge the significance of the paper. Judging a paper's significance requires a necessarily somewhat subjective projection as to where the field will move and how it will be affected by the paper under review. Despite these additional sources of noise in the value signal provided by the reviews, high-impact journals – more than specialized journals – need *precise* quality assessments if they are to realize their claim of selecting only the very best papers.

#### AUTHORS AND REVIEWERS OPERATE UNDER UNHEALTHY INCENTIVES

Even if the majority of scientists are principally motivated by a desire to find the truth and maintain a high level of personal ethics, the incentives built into the system influence the level of objectivity achieved in the writing of papers and in the evaluation process. The current system provides several unhealthy incentives:

- It rewards authors for making claims that are stronger than can be justified (as this increases the chances of selection by editors for publication in high-impact journals).
- It rewards authors for suggesting reviewers known to be friendly or supportive of the claims and for selectively citing other scientists likely to support publication (as these are more likely to be selected as reviewers).
- It rewards reviewers for spending little time reviewing (as this is time available for their own science and reviewing is not rewarded or even recorded). This encourages reviewers to decline many reviews and to avoid in-depth evaluation of the ones they accept.
- It rewards reviewers for obstructing or delaying the publications by competitors and for expediting publications by allies.

Most scientists may resist these rewards. However, an ideal system would not provide such unhealthy incentives. To obstruct or expedite publication, a reviewer need not make any false statements, but merely to gage the review's level of enthusiasm and focus on strengths or weaknesses as needed. Since the reviews and the reviewer's identity are kept secret, there is no public scrutiny of either the arguments in a review or possible conflicts of interest of the reviewer. A rogue reviewer can therefore act with impunity

and distort decisions indefinitely. The antidote to corruption is transparency – this is a central motivation for the present proposal.

### EVALUATION DELAYS PUBLICATION

The current system of journal-controlled pre-publication review delays publication of papers by months in the best case. When authors target prestigious journals, multiple rejections and rounds of review and revision, often delay publication by more than a year from the date of initial submission. Scientific papers are the major mode of formal scientific communication. Delays of many months in this crucial communication line slowdown the progress of science.

### THE SYSTEM IS CONTROLLED BY FOR-PROFIT PUBLISHERS AND INCURS EXCESSIVE COSTS

In the current system, the key function of evaluating and selecting papers is controlled by private publishing companies. Although papers are reviewed by scientists, the selection of reviewers and the decisions about publication are largely in the hands of private publishers. The publishers are professional at what they do, draw from a large amount of experience, and have a reputation to defend. However, profit maximization can be in conflict with what is best for science. The current system is immensely profitable to the publishers, so they are not natural leaders of a transition to a better system. More generally, the arguments in favor of direct public funding of not-for-profit research institutes (as opposed to buying studies from private research institutes) also apply to scientific publishing. To the extent that the free market can provide cost-efficient solutions, there is a place for the private sector. However, we need to assess whether the benefit to science of the services provided justifies the cost of the current system.

### SOME RECENT POSITIVE DEVELOPMENTS

Many positive developments in scientific publishing include the Public Library of Science (PLOS) and other open-access journals, the Frontiers journals, Faculty of 1000, and ResearchBlogging.org<sup>1</sup>. In this section we briefly describe each of these developments and explain why they represent important steps in the right direction, but do not go far enough to fully address the problems related to the way the current system utilizes peer review.

#### PUBLIC LIBRARY OF SCIENCE

The PLOS journals<sup>2</sup> combine OA with beautiful professional layout and well-designed web-interfaces. Moreover, the websites offer functionality for post-publication commentary and 1–5-star ratings on three scales (“insight,” “reliability,” and “style”) for registered users. Every paper has a “metrics” page that shows these ratings, along with usage statistics (views, pdf downloads), citation counts from multiple sources, and social-network links. The presence of these features is exemplary. PLOS ONE<sup>3</sup> takes a further step forward by using pre-publication review only to establish that a paper is “technically sound,” not to assess its importance. This is likely to render peer review more objective. The PLOS journals

combine a high scientific standards and high production value with OA.

Although the post-publication commentary and rating functionalities represent an attempt at integrating OE, these features are not widely used and thus do not yet provide a major evaluative signal at the moment. This highlights the challenge to motivate scientists to contribute to post-publication evaluation. The PLOS family of journals relies on the traditional process of secret pre-publication peer review as the core of its evaluation process. In a fully transparent post-publication system as proposed here, the editor-solicited initial reviews and ratings would be public, so every paper would have multiple reviews and ratings from this process. For specialized papers, such as those published in PLOS ONE, it is not realistic to expect many additional reviews to accumulate. Moreover, commenting on PLOS papers requires login (increasing the effort required), but comments and ratings are not part of the core evaluation mechanism (which remains secret pre-publication peer review). A scientist who might want to share an opinion has minimal motivation to use the commenting system because there is little indication that such a contribution will matter as the paper already has its mark of approval from pre-publication peer review. A signed critical comment, in particular, would mean taking a social risk without promising much positive impact. As we will see below, the change of culture required to make a transparent evaluation system work requires that the post-publication evaluations really matter as more than an add-on and that signed reviews count as mini-publications that are citable and help build the reviewer's reputation.

#### THE FRONTIERS JOURNALS

The Frontiers journals<sup>4</sup>, starting with Frontiers in Neuroscience, combine OA, a new system for constructive and interactive pre-publication peer review, web-based community tools, and post-publication quasi-democratic evaluation of papers. Moreover, Frontiers provides a hierarchy of journals from specialized (e.g., Frontiers in Computational Neuroscience) to general (Frontiers in Neuroscience). The hierarchy may be extended upward in the future.

Importantly, papers are first published in the specialized journals. Based on the additional evaluative information accumulated in the reception of the papers by the community, a subset of projects is selected for wider publication in a higher-tier journal. This has several advantages over conventional approaches: Selection for greater visibility is based on more evidence than available to traditional high-impact publications (which rely only on the few reviews and informal opinions they solicit). The higher-tier thus responds more slowly and ideally more wisely: avoiding to draw attention to findings that do not pass the test of confrontation with a larger group of peer scientists than can be asked to initially review a paper. Like PLOS, Frontiers offers web functionality for reviewing and rating, but these OE features do not yet form the core of the evaluation process.

The Frontiers system is visionary and represents a substantial step in the right direction. As for the PLOS journals, however, quality control for the lowest tier still relies on pre-publication review,

<sup>1</sup><http://researchblogging.org/>

<sup>2</sup><http://www.plos.org/>

<sup>3</sup><http://www.plosone.org/home.action>

<sup>4</sup><http://frontiersin.org/>

tolerating the evaluation inaccuracies and delays and failing to provide detailed evaluative information, such as public reviews.

### FACULTY OF 1000

Faculty of 1000<sup>5</sup> provides very brief post-publication recommendations of papers with a simple rating (“Recommended,” “Must-read,” “Exceptional”). The post-publication review idea is a step forward. However, the reviewing is limited to a select group of highly distinguished scientists – a potential source of bias. Evaluations are recommendations – there is no mechanism for negative reviews. Numerical evaluations are unidimensional thus providing only a very limited signal. Finally, the recommendation text is a brief statement, not a detailed review. In addition, the Faculty of 1000 system is a for-profit effort that is not designed or controlled by the scientific community. It is post-publication peer evaluation, but it is not OE. And it is not OA, either: The evaluations are sold by subscription.

### RESEARCHBLOGGING.ORG

ResearchBlogging.org collects blog-based responses to peer-reviewed papers. This is a big advance as it allows anyone to participate and provide evaluative information, which can be accessed through the ResearchBlogging.org website. The use of blogs is helpful in that it makes this system open. However, it also means that reviews are not permanently citable as blogs can be taken down. Moreover, as of yet the blog responses lack numerical ratings that could be automatically analyzed for paper evaluations. Blog responses are also not digitally signed for author identification, and the responses are not visible when viewing the target paper itself.

### A VISION FOR OPEN EVALUATION

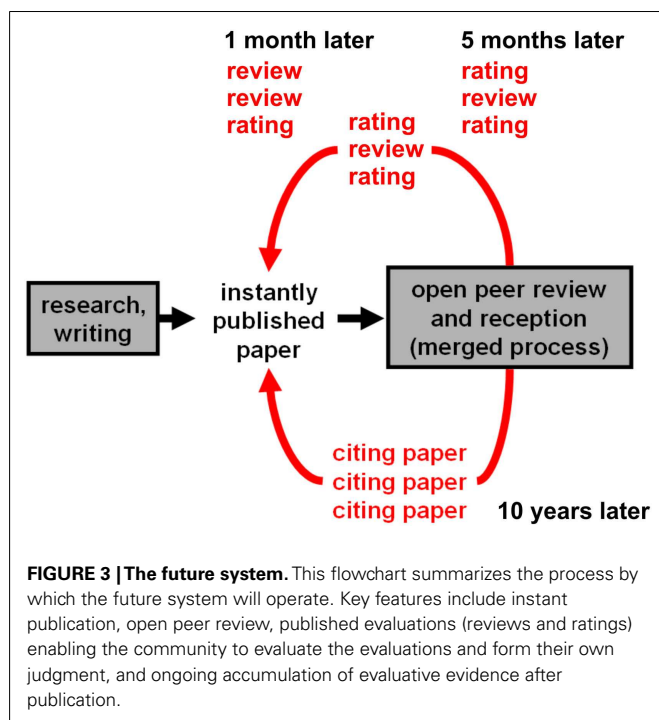
The problems of the current system can all be addressed by open post-publication peer review. The basic process of this model is summarized in **Figure 3** and illustrated in greater detail in **Figure 4**.

#### OPEN

Any scientist can instantly publish a peer review on any published paper. The scientist will submit the review to a public repository (see also Florian, 2012 in this collection). Reviews can include written text, Figures, and numerical quality ratings on multiple scales. The repository will link each paper to all its reviews, such that readers can readily access the evaluative meta-information whenever they view a paper. Peer review is open in both directions: (1) Any scientist can freely submit a review on any paper. (2) Anyone can freely access any review as soon as it is posted.

#### POST-PUBLICATION

Evaluations are posted after publication, because a paper needs to be publicly accessible in order for any scientist to be able to review it. Post-publication reviews can add evaluative information to papers published in the current system (which have already been secretly reviewed before publication). For example, a controversial



paper appearing in *Science* may motivate a number of supportive and critical post-publication reviews. The overall evaluation from these public reviews will affect the attention given to the paper by potential readers. The written reviews may help readers better understand and judge the paper.

#### PEER REVIEWS

Like the current system of pre-publication evaluation, the new system relies on peer reviews and ratings. For all of its faults, peer review is the best mechanism available for evaluation of scientific papers. Note however, that public post-publication reviews differ in two crucial respects:

- (1) They do not decide about publication – as the papers reviewed are already published.
- (2) They are public communications to the community at large, not secret communications to editors and authors.

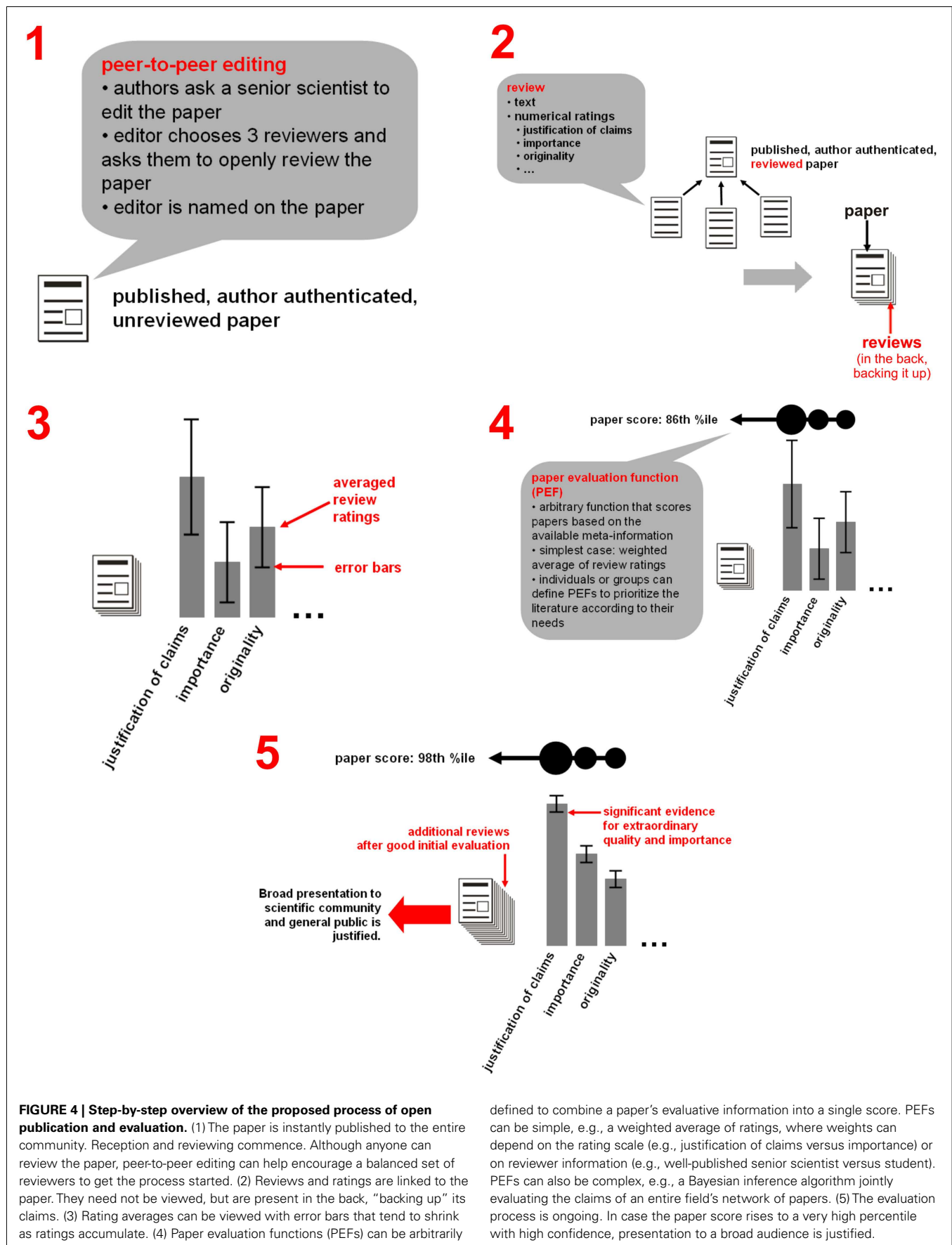
This makes the peer review more similar to getting up to comment on a talk presented at a conference. Because the reviews are transparent and do not decide about publication, they are less affected by politics. Because they are communications to the community, their power depends on how compelling their arguments are to the community. This is in contrast to secret peer review, where unconvincing arguments can prevent publication because editors largely rely on reviewers' judgments and reviewers are not acting publicly before the eyes of the community.

#### PEER RATINGS

The term “evaluation” refers to both reviews and ratings. Like peer reviews, peer ratings are used by many journals in the current system. However, the valuable multi-dimensional quantitative

<sup>5</sup><http://www.facultyof1000.com/>







information they provide remain secret. The OE system will enable explicit ratings on multiple scales that reflect both the confidence that the claims are veridical and the importance of the paper. Scales will include “justification of claims,” “novelty of claims,” and “significance of claims.” The system will also include a simple syntax for freely introducing new scales within any evaluation. All this requires is to give the new scale a name that clearly denotes its meaning and to provide a rating.

### MULTIPLE LENSES ONTO THE LITERATURE

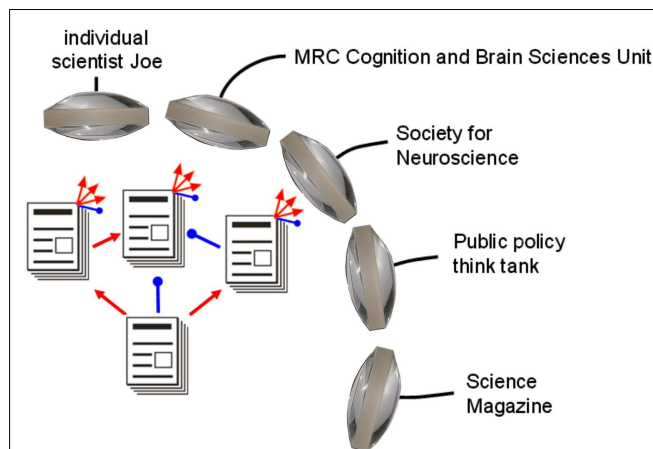
The necessary selection of papers for reading can be based on the reviews and their associated numerical judgments. Any group or individual can define a PEF based on content and quality criteria. A PEF could for example, rely only on signed ratings from post-PhD scientists, and weight different rating scales in a particular way. A PEF could also employ social-network information, e.g., to down-weight ratings from reviewers that are associated with the authors. Social networks could also contribute evaluative information on papers to PEFs, including usage and sharing statistics as well as ratings (Lee, 2012; Priem and Hemminger, 2012; Walther and van den Bosch, 2012; Zimmermann et al., 2012; all in this collection). Beyond weighted averaging, PEFs could use complex recurrent inference algorithms, e.g., to infer probabilities for the title claims of papers. Social web and collaborative filtering algorithms (Goldberg, 1992; Breese et al., 1998; Schafer et al., 2007) will be applied to this problem. However, evaluating the scientific literature poses unique challenges and requires greater transparency and justification than product recommendation systems. The development of PEFs will build on and extend the existing literature on collaborative filtering systems. There will be a plurality of PEFs prioritizing the literature from multiple perspectives (Figure 5). When reviewers start using a new rating scale in their evaluations, PEFs may utilize the ratings on the new scale if the evaluative evidence the scale provides is thought to justify its inclusion.

### WEB-PORTALS AS ENTRY POINTS TO THE LITERATURE

Web-portals (“subject focal points,” Smith, 1999) will serve as entry points to the literature, analyzing the numerical judgments in the reviews by different criteria of quality and content (including the use of meta-information about the scientists that submitted the reviews). There will be many competing definitions of quality – a unique one for each web-portal and each individual defining his or her own PEF. Web-portals can define PEFs for subcommunities – for scientists too busy to define their own. A web-portal can be established cheaply by individuals or groups whose members share a common set of criteria for paper prioritization.

### MERGING REVIEW AND RECEPTION

Currently review is a time-limited pre-publication process and reception of a paper by the community occurs later and over a much longer period, providing a very delayed – but ultimately important – evaluative signal: the number of citations a paper receives. Open post-publication peer review will remove the artificial and unnecessary separation of review and reception. It will provide for a single integrated process of open-ended reception and review of each paper by the community. Important papers will accumulate a solid set of evaluations and bubble up in the process – some of them rapidly others after years.



**FIGURE 5 | A plurality of paper evaluation functions (PEFs) provides multiple lenses onto the literature.** Organizations and individuals can define PEFs according to their own priorities and make the resulting paper rankings publicly available. Competing PEFs provide multiple perspectives. Moreover, the OE system becomes “ungameable” as PEFs respond to any attempts by individual scientists or groups to take advantage of weaknesses of current PEFs. With constantly evolving PEFs, each scientist and organization is motivated to aim for truth and objectivity. Red and blue pointers correspond to “excitatory” and “inhibitory” evaluative links, which could be represented by positive and negative numerical ratings. Beyond simple averaging of ratings, PEFs could employ sophisticated inference algorithms to jointly estimate the probabilities of all papers’ title claims.

### SIGNED AND ANONYMOUS EVALUATIONS

There is some evidence that the threat of revealing the reviewer’s identity to the authors (van Rooyen et al., 1999) or of making a review public (van Rooyen et al., 2010) may just deter reviewers and do little to improve review quality. This highlights the need to give reviewers a choice of whether or not to sign. Moreover, defining reviews as open letters and mini-publications will create a different culture, in which scientists define themselves not only through their own work, but also through others’ work they value. Signed evaluations have the advantage that they attach the reviewer’s reputation to the judgment, thus alleviating abuse of reviewer power (Walsh et al., 2000; Groves, 2010). Anonymous reviews have the advantage that they enable reviewers to criticize without fear of negative consequences (Khan, 2010). Both types are needed, and a scientist will make this choice on a case-by-case basis. The anonymous option will encourage communication of critical arguments. But to the extent that an argument is objective, sound, and original, a scientist will be tempted to sign in order to take credit for his or her contribution. In analyzing the review information to rank papers, signed reviews can be given greater weight if there is evidence that they are more reliable. Reviewers can digitally sign their reviews by public-key cryptography<sup>6</sup>. The idea of digitally signed public reviews has been developed here<sup>7</sup>, where further discussion and a basic software tool that implements this function can be found.

<sup>6</sup>[http://en.wikipedia.org/wiki/Public-key\\_cryptography](http://en.wikipedia.org/wiki/Public-key_cryptography)

<sup>7</sup><http://code.google.com/p/gpeerreview/>

## REVIEWS AS OPEN LETTERS TO THE COMMUNITY

Reviews will no longer be secret communications deciding about publication. They will be open letters to the community with numerical quality ratings that will influence a paper's visibility. OE will initially build on the current system by providing higher-quality transparent evaluations of papers that have already been reviewed secretly before publication. As long a traditional peer review is in place, we expect mainly important papers to attract additional OEs. The original pre-publication reviewers could use the OE system to make their reviews (updated to reflect the published revision) public, so that their work in reviewing the paper can be of benefit to the readers of the paper and to the community at large.

## IMPROVING EVALUATION QUALITY

The quality of the evaluative signals will be improved by post-publication review for a number of reasons:

- (1) Since reviews are open letters to the community, their power is dependent on how compelling they are to the community. (In the present system, rejecting a paper does not require an argument that would hold up under the scrutiny of the community. For a high-impact journal, for example, all it takes is to say that the paper is good, but not sufficiently surprising.)
- (2) The system will include signed evaluations, so the reviewer's reputation is on the line: he or she will want to look objective and reasonable. (Anonymous evaluations can be down-weighted in assessment functions to the degree that they are thought to be unreliable.)
- (3) Important papers will accumulate more evaluations (both reviews and ratings) over time as the review phase is open ended, thus providing an increasingly reliable evaluative signal.

Ratings, like reviews, can be signed and will enable us to help steer the attention of our field without investing the time required for a full review. Early signed ratings that turn out to be solid can contribute to a scientist's reputation just as reviews can. As researchers read and discuss the literature in journal clubs around the world as needed for their own research, the expert judgments are already being performed behind closed doors. The OE system will provide a mechanism for feedback of this valuable information into the public domain. With PEFs in place to summarize the evaluations, journal prestige will eventually not be needed anymore as an evaluative signal.

## COMMUNITY CONTROL OF THE CRITICAL FUNCTION OF PAPER EVALUATION

Open evaluation means that the scientific community organizes the evaluation of papers independently, thus taking control of this critical function, which is currently administered by publishers. Evaluation is the key function that currently keeps science dependent on for-profit publishers. Achieving OE, therefore, will also help accelerate the ongoing shift toward general OA. Conversely, OA is a requirement for true OE, as only openly accessible papers can be evaluated by the entire community. OA and OE are the two complementary pieces of the ongoing paradigm shift in scientific publishing.

## A DIVISION OF POWERS BETWEEN THE ACCUMULATION OF EVALUATIONS AND THE PRIORITIZATION OF THE LITERATURE

A core feature of this proposal is a clear division of powers between the OE system, which accumulates reviews and ratings and links them to the papers they refer to, and the PEFs, which combine the evaluative evidence so as to prioritize the literature from particular perspectives. This division of powers requires that the evidence accumulated by the OE system is publicly available, so that independent groups and individuals can analyze it and provide PEFs. This division of powers ensures transparency and enables unrelated groups and individuals to freely contribute to the evaluative evidence and to its combination for prioritizing papers. For example, if a group of scientists started doing mutual favors by positively evaluating each other's papers, an independent group could build a PEF that uses only signed evaluations and downweights evaluations from individuals within cliques of positive mutual evaluation. Conversely, when a web-portal claims to combine the evaluative evidence by a given PEF to compute its paper ranking, anyone can re-implement that algorithm, run it on the public evaluative evidence, and check the ranking for correctness. This fosters a culture in which we keep each other honest, and in which public interest and self-interest are aligned. When the process is entirely transparent and competing PEFs evolve in response to any attempts to game the system, an individual's best bet is to act according to the criteria of objectivity he or she believes will eventually prevail.

## A SPECIFIC PLAN FOR A MINIMALIST OPEN EVALUATION SYSTEM

What are the minimal requirements for a web-based OE system for accumulating evaluations? We would like the system to enable rapid ratings, signed or unsigned, and also multi-dimensional ratings and in-depth reviews. A key consideration is the time it takes for users to provide ratings as this will determine the efficiency of the system and, thus, the volume of evaluative evidence accumulated. I will now describe a prototype that meets minimum requirements and is designed to "seduce" the user to provide more detailed information.

## WHAT KIND OF RATING SCALE?

The quickest rating is clicking a "like" button. While this has proven useful for prioritizing items in non-scientific web systems, it is not ideal for evaluating scientific papers. The key argument against one-click ratings is that they provide continuous valuations only in aggregate. Counting the number of likes confounds the amount of exposure a given item (e.g., a paper) has received (how many people considered clicking "like") with the value attributed to it. Adding a "dislike" button enables us to consider the balance of likes and dislikes. However, a continuous valuation requires a sizeable number of contributions, and error bars on the valuation require even more contributions. "Like" and "dislike" buttons, therefore, are ideal for sampling casual judgments of large numbers of people, but less suited for our present purpose, i.e., sampling careful judgments of small numbers of people.

I therefore suggest using an overall rating scale as the first evaluative piece of information. The fastest way to collect a continuous judgment might be a click on a continuous scale on the screen.

However, we are interested in careful deliberate evaluations. We therefore prefer the user to decide on a numerical rating. A numerical rating is also better suited for being explicitly remembered and communicated. Entering one number takes only a little longer than a click.

The next question is how the single scale should be defined. Rating scales for movies and other cultural items sometimes use a five-star system. However, a five-level scale appears too coarse to reflect individual scientists' quality judgments on papers and also does not provide a sufficiently fine-grained signal for prioritization entire literatures. A higher resolution appears desirable, e.g., a number between 0 and 100. Bounding the ratings between a lowest and highest value provides an intuitive definition of its units, e.g., from worst to best imaginable. Ideally, however, the units of the scale should be defined more precisely than by a mere specification of bounds. In that case bounding the scale is not necessary.

A rating could be conceptualized as a "weight," which the rater thinks should be given to the paper in combining the evidence on a scientific question (as in optimal linear estimation). This would suggest that 0 should be the lowest possible rating. A rating of 0 would communicate the judgment that the paper's contents are best ignored in order to arrive at the truth. Note, however, that limiting the scale to positive values entails that the average across multiple noisy ratings will be positively biased (i.e., the average will always be greater than 0 even if the paper deserves a weight of 0). To address this shortcoming, ratings could comprise negative as well as positive numbers. This possibility is illustrated in **Figure 5**. Positive and negative ratings could provide excitatory and inhibitory connections in an evaluative network. This would enable negative judgments to balance positive judgments and reduce the effective weight given to a paper (as estimated by an average across the ratings) all the way to 0. In addition, it might be desirable to collect a confidence rating in addition to the rating itself. With a confidence range, the rater could communicate not just a point estimate but a full probability density over ratings reflecting subjective certainty. This would be useful for Bayesian inference on the basis of multiple ratings. Such an inference procedure could also include a probabilistic model of each rater's reliability (e.g., based on past performance). Although negative ratings and confidence ranges will likely prove useful for some of the scales that will come to be used in the system, the first and overall scale for the minimalist system we describe in this section is restricted to positive values, as this is more consistent with this scale's content and function, as explained below.

Beyond the resolution and range of the scale, we need to decide the content: What evaluation criteria should be captured by the first scale (for which we expect to accumulate the greatest number of ratings)? The scale's definition must be highly general as any specific choice we make is going to be problematic. Say we defined the scale as measuring the "justification of the claims" of the paper. A user might find a technical paper that is highly justified in its claims less significant than a bold paper that presents a ground-breaking theory and still makes a reasonable case for its claims. Other users will have different priorities. While the proposed system ultimately addresses this issue by enabling multi-dimensional ratings (including open-ended definition of new scales), it still faces a decision for the first scale.

## DESIRED-IMPACT RATINGS

We must not put the user in a double bind, where the scale is defined by one criterion, but he or she would prefer to judge by another, in awareness of the real-world consequences of the judgment on the visibility and thus ultimately on the impact of the paper. I therefore propose that the single overall scale should be the "desired impact" for the paper. This describes the actual effect the scale is meant to control and thus avoids the double bind. A user who feels that justification of claims should be the most important criterion will judge desired impact by this criterion. A different user might give more weight to the originality of the ideas put forward. Defining the scale as "desired impact" acknowledges the inherent subjectivity of judging the significance of scientific papers.

Note that the proposed overall scale of "desired impact" is not the only scale that should be used. Other scales will focus explicitly on the justification of the claims of a paper and on other specific evaluative dimensions. Note also that "desired-impact" ratings express *desired*, not *predicted* impact. One might predict great impact for a paper one considers incorrect. But most of us would not *desire* high impact for such a paper. The desired-impact rating enables scientists to judge by their own criteria (including veracity and importance) what impact a paper deserves.

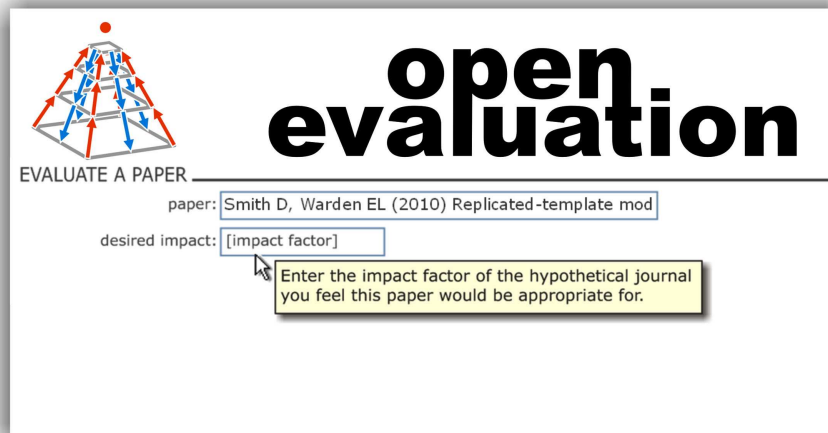
The next question is how desired impact should be expressed numerically. I propose the use of a unit that scientists already understand: the IF. IFs are used in the current system of scientific publishing for evaluating journals. Journal IFs are problematic, especially when they are misinterpreted as measures of the quality of the papers published in a given journal. However, they are widely understood and grounded in the citation success of papers. The IF of a journal is the average number of citations in the present year received by papers published by the journal in the previous 2 years. We can loosely interpret the IF as the average citation success of a paper in the 2 years following the year of its publication.

We define the first scale as "desired impact" in IF units. The IF unit is redefined to apply to a particular paper as measuring the number of citations the paper should receive in the 2 years following the year of its publication, so as to be considered by the user as having received an appropriate amount of attention. Alternatively, we can think of the desired-impact rating as the IF of the hypothetical journal that the paper is deemed appropriate for.

## RAPID RATING, OPTIONAL IN-DEPTH REVIEWING

**Figure 6** presents a web-interface that provides the functionality for rapidly collecting desired-impact ratings, while "tempting" the user to provide more detailed evaluative evidence. First, the user specifies the paper to be evaluated. This can be done either by clicking a link in PubMed, Google Scholar, or a similar search engine, or by explicitly specifying the paper in the OE interface. The user then enters the desired-impact rating, whereupon a "Submit unsigned evaluation" button appears along with a new field for optional signing of the evaluation. The user can click "Submit unsigned evaluation" and be done in about 20 s, or sign, which might require an additional 10 s.

Signing can utilize existing web identification and authentication technology. It could be automatized using active logins in scientific or non-scientific social networks. For example, Google Gmail, facebook, and Apple iTunes all use such technology. But

**Step 1 (rate, 20 seconds)**


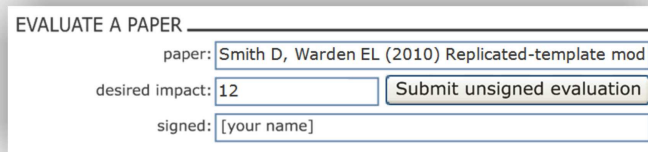
**open evaluation**

EVALUATE A PAPER

paper:

desired impact:

Enter the impact factor of the hypothetical journal you feel this paper would be appropriate for.

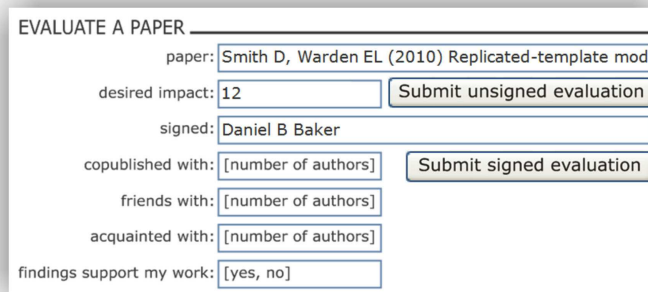
**Step 2 (sign, 10 seconds)**


EVALUATE A PAPER

paper:

desired impact:

signed:

**Step 3 (disclose, 30 seconds)**


EVALUATE A PAPER

paper:

desired impact:

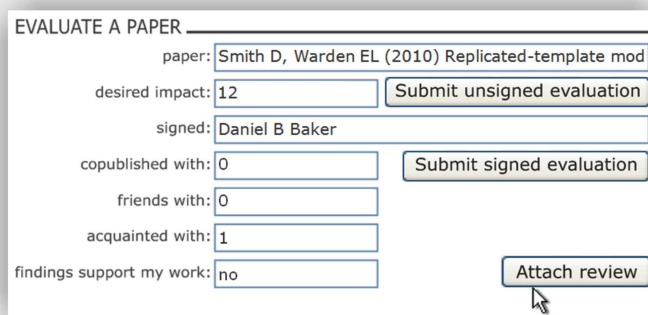
signed:

copublished with:

friends with:

acquainted with:

findings support my work:

**Step 4 (review, open ended)**


EVALUATE A PAPER

paper:

desired impact:

signed:

copublished with:

friends with:

acquainted with:

findings support my work:

**FIGURE 6 | A minimalist system for accumulating evaluations (ratings and reviews).** Steps 1–4 illustrate a user's interaction with an envisioned web interface. (1) Rate: The user selects a paper to evaluate (either by clicking on an evaluation link associated with the paper, or by specifying the paper in the top entry field. The user then enters a single overall numerical rating (desired impact in impact-factor units), whereupon a button labeled "Submit unsigned evaluation" appears (shown in step 2). By clicking this button, the user can submit the overall rating anonymously and terminate the process with a total time-investment of about 20 s. (2) Sign: alternatively, the user can choose to sign the evaluation by entering

his or her name, whereupon a button labeled "Submit signed evaluation" appears. By clicking this button, the user can submit the overall rating as a signed evaluation with a total time-investment of about 30 s. (3) Disclose: optionally the user can disclose information on social links to the authors and personal stake in the claims before submission, which might take another 30 s. (4) Review: finally, the user can attach a written review (a txt, doc, or pdf), which can include detailed ratings on multiple scales (in a standard syntax that makes the ratings extractable and enables open-ended definition of new scales), as well as written arguments and figures.

even if the scientist just signed by name in a text field, the system could work, because all evaluations are public, and identity theft in OE could be righted retrospectively.

The motivation to sign would come from the greater weight certain PEFs will assign to signed evaluations. In some of these PEFs this weight will also depend on an evaluation of the signing scientist. In addition, signing evaluations contributes to the reputation and visibility of the scientist.

After signing, the user has the option to disclose information about social links to the authors and about any personal stake in the results of paper. Within another 30 s, the user can disclose how many of the authors he (1) has co-published with in the past, (2) is friends with, and (3) is acquainted with, and (4) whether the findings reflect positively upon his or her own work. These ratings are made in an honor system. However, since they are public information that can be verified, there is a strong disincentive to misrepresent potential conflicts of interest. As for signing, the positive motivation for disclosing comes from the greater weight some PEFs will assign to ratings, for which this information is available.

Finally, the user is given the option to attach a review. The review can be attached in a suitable format for being read by people and analyzed by PEFs. The existing formats txt, doc, or pdf could initially serve this purpose, although more structured and flexible formats might come to be preferred. A review can contain ratings on multiple scales (which are labeled in a flexible syntax that enables the user to introduce additional scales as needed to capture the quality of the work), along with text and figures. Such a review is an instant citable, mini-publication, providing added motivation for contributing to the process.

### A SPECIFIC PLAN FOR A MINIMALIST PAPER EVALUATION FUNCTION

The web-based OE system we described above can accumulate the evaluative evidence. However, the evidence still needs to be

combined for prioritizing the literature. We have stressed the need for a division of powers between these two components of the evaluation process, and for a plurality of perspectives on the literature in the form of multiple competing PEFs. To make the concept of a PEF more concrete, I propose a blueprint for a general-purpose PEF called “sciture” (Figure 7).

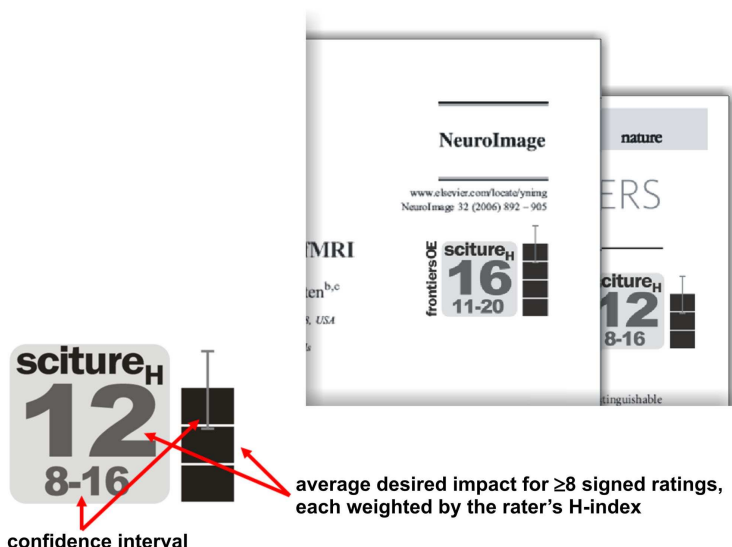
Sciture stands for “scientific citation future.” This particular PEF uses only the desired-impact scale, enabling it to draw from a larger number of ratings than PEFs that combine multiple rating scales (for which we expect to accumulate fewer scientists’ ratings). A paper’s sciture is the impact projected for the paper by the scientists that rated the paper. The sciture is the desired impact in impact-factor units averaged across the scientists who signed their ratings.

There are two variants of the index. The first (simply called sciture) uses an unweighted average of all signed ratings, so as to give raters equal influence. The second is called  $sciture_H$  and weights each rating by the rater’s  $H$ -index (Hirsch, 2005), thus giving more weight to scientists whose own publications have had a greater impact. Note that this index excludes laypeople and young scientists who have never published or whose publications have never been cited. This may be seen as an advantage or a disadvantage and may motivate the definition of alternative PEFs. Note also that the present definition of sciture ignores conflict of interest information. This could be changed in case there were evidence that positive evaluation cliques distort the ratings.

### THE ULTIMATE GOAL: FREE, INSTANT, OPEN-ACCESS PUBLISHING, PEER-TO-PEER EDITING, AND OPEN EVALUATION

#### FREE INSTANT PUBLISHING

Once OE provides the critical evaluation function, papers themselves will no longer strictly need journals in order to become part of the scientific literature. They can be published like the



**FIGURE 7 | A minimalist paper evaluation function.** As one possible general-purpose PEF, I suggest the “sciture<sub>H</sub>” index, which is an average of at least eight scientists’ desired-impact ratings (in impact-factor units), weighted by the scientists’  $H$ -indices. Such an index could serve to provide ongoing

open evaluation of papers published under the current system. An icon summarizing the index and its precision (left) could be added to online representations of papers (right), either by the publishers themselves or by independent web-portals providing access to the literature.



reviews: as digitally signed documents that are instantly publicly available. OE will provide evaluative information for any sufficiently important publication. With OE in place, there is no strong argument for pre-publication review. The binary decision for or against publication will be replaced by graded evaluative evidence, that is summarized by PEFs. Publication on the internet can, thus, be instant and reviews will follow as part of the integrated post-publication process of reception and evaluation.

### PEER-TO-PEER EDITING

Peer-to-peer editing can help to get the evaluation process started and to ensure that the initial two to four reviewers are somewhat balanced in terms of biases and expertise. Balance is particularly important in the initial phase, because a small number of negatively biased initial reviews can nip a paper's OE process in the bud. After publication, the author asks a senior scientist in his or her field to serve as *editor* for the paper. If the senior scientist accepts, an acknowledgment of his or her role as editor will be added to the paper. The editor's job is to select two to four reviewers and to email them (via an automatic system that uses standard invite texts) with the request to publicly review the paper. If they decline, the editor is to find replacements. However, anyone else is allowed to review the paper as well. In particular, the author may also inform other scientists of the publication and ask them to review the paper. Author- and editor-requested reviews will be marked as such. Reviewers could be asked to state whether or not the review has been requested by an editor or suggested (explicitly or implicitly) by the authors. Requested as well as unrequested reviews can be signed or unsigned. Editors must not have been at the same institution or on any paper with the authors. Reciprocal or within-clique editing is monitored and discouraged. Since the editors are named, PEFs can detect within-clique editing and reviewing and downweight within-clique reviews and within-clique-editor-requested reviews.

### REVISIONS

If the weight of the criticism in the accumulated reviews and the importance of the paper justify it, the authors have the option to revise their paper. The revision will then be the first thing the reader sees when accessing the paper and the authors' response to the reviews may render the criticism obsolete. However, the history of revisions of the paper, starting from the original publication will remain accessible in perpetuity.

### REVIEWS AS MINI-PUBLICATIONS

Reviews will no longer be secret communications deciding about publication. They will be open letters to the community with numerical quality ratings that will influence a paper's visibility on web-portals. The quality and quantity of signed reviews written by a given scientist will be one of the determinants of his or her status. This will greatly enhance the motivation to participate in the evaluation process. With a general OE system in place, reviewing activity can be analyzed with the same methods used to analyze other publication activity. **Figure 8** contrasts the nature of a review in the current and in the proposed system. Through transparency, the proposed system replaces the unhealthy incentives of the current system (i.e., to minimize the time spent reviewing and to exert

#### Current

- secret communication to authors and editors
- decides about publication
- reviewer's motivation
  - selfless: scientific objectivity
  - selfish: science politics
- a weak argument can kill a paper

#### Future

- open letter to the community
- evaluates published work
- reviewer's motivation
  - selfless: scientific objectivity
  - selfish: looking smart and objective in public
- an argument is as powerful as it is compelling

**FIGURE 8 | The nature of a review in the current and future systems.**

This juxtaposition of the key features of a review in the current and the future system points to some essential changes in scientific culture that the transition will entail.

political influence) by healthy ones (i.e., to contribute objective and reasonable evaluations so as to build one's reputation).

### A DIFFERENT CULTURE OF SCIENCE

Open evaluation goes hand in hand with a new culture of science. This culture will be more open, transparent, and community controlled than the current one. We will define ourselves as scientists not only by our primary research papers, but also by our signed reviews, and by the prior work we value through our public signed ratings. The current clear distinction between the two senses of "review" (as an evaluation of a particular paper and as a summary and reflection upon a set of prior papers) will blur. Reviews will be the meta-publications that evaluate and integrate the literature and enable us as a community to form coherent views and overviews of exploding and increasingly specialized literatures. Evaluation of scientific work and distillation of the key insights are at the heart of the entire endeavor of science. The scientific community will therefore take on the challenge of designing and continually improving the evaluation system. This includes design of the human-computer interfaces, design of the web-mediated interactions between humans, and design of artificial-intelligence components that will help evaluate and integrate our insights. Designing the OE system will lead us to the ultimate challenge: to design the collective cognitive process by which science, globally connected through the web, constructs our view of the world, and ourselves.

### DISCUSSION

The discussion is structured by critical questions that I have encountered when discussing this proposal.

#### IF PEER REVIEW OCCURS ONLY POST-PUBLICATION, WON'T THE LITERATURE BE SWAMPED WITH LOW-QUALITY PAPERS THAT ARE NEVER EVALUATED?

Yes, but that's not a problem. Peer review currently serves as a barrier to entry into the literature, serving to maintain a certain quality standard. Removing this barrier might seem dangerous in that it might open the gates to a flood of low-quality papers. In other notable proposals of public peer review, pre-publication

review therefore still plays a role (e.g., Bachmann, 2011; Kravitz and Baker, 2011; Pöschl, 2012; Sandewall, 2012; all in this collection). Here, we argue that pre-publication peer review is not needed or desirable. Peer evaluation will help us select what to read, but it will not prevent us from reading papers that have not (yet) been evaluated.

### ***Minimal formal barriers to publication***

For a paper to become a citable and permanently archived publication, the authors' identities need to be verified. In addition, a restriction could be placed on the volume of work per author (e.g., 12 papers per year). This would help prevent computer-generated content from being submitted. Beyond these formal restrictions, authors will be aware that low-quality publications will damage their reputations. Scientific papers require minimal storage (compared to other cultural products, such as movies) and their number is small per capita of the population and year. Although the total storage required will be substantial, our technology can handle it.

### ***Only published papers can be publicly evaluated***

Peer evaluation cannot be truly open (i.e., public) unless the paper is publicly available (i.e., "published"). A public peer review, thus, is post-publication by definition. A pre-publication stage would be merely a matter of labeling published papers as either "under review" or "reviewed" (i.e., "properly published"). However, OE is to be ongoing and incremental, and the evaluative signal continuous and multidimensional. Labeling already published papers as "reviewed" or "properly published" at some stage merely amounts to imposing an arbitrary threshold on some PEF. There is no clear motivation, thus, for dividing OE into two stages.

### ***The twilight zone of unevaluated papers***

Some published papers will never get a single review or rating; this is not a problem. There will be a new twilight zone of published, citable, but unevaluated papers. As readers, we do not mind this, because twilight papers will not come to our attention unless we explicitly search for them. As authors whose work remains in the twilight, we will learn that we need to connect better with peers through conferences, conversations, and high-quality work, to earn enough respect to find an initial audience, and a peer-to-peer editor. In case we are too far ahead of our peers to be understood, our twilight publications might be discovered later on. The future system will thus provide a mechanism for publication of science that defies the dominant scientific paradigm, is unpopular for other reasons, or simply difficult to understand. However, there is no instant mechanism for distinguishing the bad from the brilliant, but misunderstood. It is therefore necessary to provide permanent access to both, and unavoidable that a proportion of the literature will receive little attention and no proper evaluation.

### **WHAT IF THERE ARE TOO FEW EVALUATIONS FOR A PAPER?**

#### ***Papers with less than eight ratings will come with large error bars, or without error bars***

Many papers will receive some evaluations, but not enough for reliable averages. These papers are under evaluated as are all papers in the current system. In the proposed system, however, the lack of reliable evaluation will be reflected in the absence (or large range) of the error bars on the overall score from a given PEF.

### ***Important papers will be broadly and deeply evaluated***

Important work will eventually be read, rated, and reviewed. Because a scientist's time is a limited resource, broad and deep evaluation can only be achieved for a subset of papers. Broad evaluation means that many scientists from different fields participate in the evaluation. Deep evaluation means that experts in the field provide in-depth evaluations and commentary on the details. To the extent that an initial set of reviews brings more attention to a paper, it will tend to be more broadly and deeply evaluated. This selective and recurrent allocation of the field's attention is a key feature of the proposed system. Selective recurrent rating and reviewing ensures that we have a reliable evaluation before raising a paper to global visibility within science and before bringing it to the attention of the general public.

### **HOW CAN SCIENTISTS BE MOTIVATED TO SUBMIT REVIEWS IN AN OPEN PEER-REVIEW SYSTEM?**

#### ***Scientists accept requests to review papers in the current system – this will not change***

In the current system, scientists are approached by editors and asked to review new papers. They regularly comply. In the new system, they will be approached similarly often through peer-to-peer editing with the same request – only the reviews will be public. There is some evidence that potential reviewers are more likely to decline to review when they are told that their name will be revealed to the authors (van Rooyen et al., 1999) or that their review might be publicly posted (van Rooyen et al., 2010). This reflects the culture of the current system, in which the reviewer expects no benefits, except the opportunity to read new work, to help improve it, and to contribute to the publication decision. In this context, removing anonymity appears to have no upside and could pose a threat. In the future system, however, reviews will be mini-publications that bring substantial benefits to the reviewer.

#### ***The motivation to review a paper is greater if the review is an open letter to the community***

The fact that reviews are public makes reviewing a more meaningful and motivating activity. In terms of power, the reviewer loses and gains in the transition to the proposed system: The reviewer loses the power to prevent or promote the publication of a paper by means of a secret review. The reviewer gains the power to speak to the whole community about the merits and shortcomings of the paper, thus building his or her reputation. The power lost is the secretive and political kind of power, which corrupts. The power gained is the open and objective kind of power that motivates constructive critical argument.

#### ***Signed reviews will be citable mini-publications contributing to a scientist's reputation***

Reviews will be citable publications in their own right. This will motivate reviewers in terms of quality and quantity. Moreover, reviews can themselves be subject to second-order peer evaluation. Reviewing will gain in importance, because it is critical to the hierarchical organization of an exploding body of knowledge. Reviewing will therefore become a scientific activity that is more publicly valued and formally acknowledged than it currently is.

Conversely, the absence of a contribution to OE will reflect negatively on a scientist. These factors will increase the motivation to participate in the evaluation process.

### WILL SIGNED REVIEWS NOT BE POSITIVELY BIASED?

Signed reviews might indeed be affected by a positive bias (Walsh et al., 2000). Reviewers might want to please particular authors or groups (specific bias), or they might want to be perceived as nice people (general bias). However, this is not a problem for four reasons:

- (1) *Reviewers are motivated to minimize bias when they sign their reviews: their reputation is on the line:* The perception of a specific bias attributable to the reviewer's academic or social connections or to a self-serving preference for certain theories would seriously threaten the reviewer's reputation. To a lesser extent, a reviewer who signs will also be motivated to minimize general positive bias, which might result from the desire to appear to be a nice person. A general "niceness bias" would suggest that the reviewer is undiscerning and thus fails to contribute critical judgment.
- (2) *A general positive bias will not compromise the assessment of the relative merit of different papers:* Even if each reviewer were affected by niceness bias to some degree, the relative merit of different papers could still be judged. The extreme scenario would be an endorsement culture, where only positive reviews are ever signed. This is comparable to reference letters, which are meant to help evaluate people's abilities. Reference letters are affected by massive positive biases, but still serve their purpose. Even if all signed evaluations were positive endorsements, the number of endorsements, the numerical ratings, and the level of enthusiasm of the positive reviews would still offer valuable measures of the community's appreciation of a paper.
- (3) *Biases of signing reviewers can be measured and corrected for by PEFs:* For a given reviewer, the set of signed reviews written and the distribution of signed numerical ratings given are public information. PEFs could therefore estimate and remove biases. For example, each reviewer's ratings could be converted to percentiles, reflecting the relative rating in comparison to the other studies reviewed by the same person. In addition, a reviewer's general bias could be assessed by comparing each of his or her ratings to the mean of the other reviewer ratings across all papers reviewed. As for specific biases reflecting academic or social connections or preferences for particular theories, these too could be automatically assessed. The suggested minimalist OE system already includes optional disclosure of information that might suggest biases (i.e., collaborative or social connections to the authors and a personal stake in the results). When the reviewer does not volunteer such information, his or her ratings could be downweighted preemptively. Moreover, analyses of social and academic networks and of published papers could be used to estimate the probability of a conflict of interest. Again, PEFs could use such estimates to adjust the weight assigned to a reviewer's ratings.
- (4) *Signing reviews is optional:* If a reviewer feels timid about signing a critical argument, he or she can contribute the argument without signing the review. Unsigned reviews and ratings might be given less weight in some PEFs. However, other reviewers who

invest enough time in the paper to read the previous reviews may pick up the argument if it is compelling in their own signed reviews. Note that there is no need for an ethical requirement that a single scientist either sign or not sign all reviews. Instead signed and unsigned reviews serve complementary positive roles and the choice between them motivates a richer and freer exchange of arguments.

### HOW CAN REVIEWS AND REVIEWERS BE EVALUATED?

A key decision in the design of a PEF is how to weight the ratings of different reviewers. First, signed ratings can be weighted by evaluations of the reviewers who gave them. In the sciture<sub>H</sub> index suggested above, each rating is weighted by the reviewer's *H*-index. Alternative indices of the reviewer's general scientific performance could equally be used (Kreiman and Maunsell, 2011 in this collection). However, it might be preferable to evaluate a reviewer's performance at the specific task of reviewing, e.g., by estimating the predictive power of their past reviews or by relying on meta-reviews of their past reviews (see Wicherts et al., 2012; in this collection). Second, without evaluating the reviewer, we can directly evaluate a given rating or review to determine its weight. Some ways of evaluating a particular review or rating and the overall performance of a reviewer are as follows.

#### **Reviews and ratings can be evaluated by meta-reviews and meta-ratings**

A review is a mini-publication that evaluates another publication. That other publication can be another review. This simple mechanism enables scientists to rate and review ratings and reviews. It can also serve as a mechanism for authors to respond to reviews. PEFs exploiting meta-ratings can recursively compute the weights, employing heuristics that prevent meta-raters from neutralizing substantial judgments. For example, a PEF might ignore unsigned meta-ratings and meta-ratings signed by one of the authors of the original paper.

#### **Reviews can be evaluated through reviewer self-report of relevant information**

Reviewers can self-report numerical information relevant to weighting their reviews. This information would be part of the ratings block in the review text. In the minimalist OE system described above, reviewers can disclose personal links to the authors of the paper and a personal stake in the claims. In addition, reviewers could self-report a confidence interval for their ratings. Self-report of confidence would enable optimal statistical combination of multiple reviewers' contributions in PEFs. Reviewers would have an incentive to accurately assess their own confidence because an error with high self-reported confidence would have a stronger impact on their reputation. Another potentially helpful piece of information is a reviewer's time-investment in the review. A judgment based on several days of reading the paper, thinking about it, and further researching key issues might be given greater weight than a judgment made in passing. Self-report of time-investment would be an honor system. However, time-investment ratings could be summed to check a reviewer's total claimed time-investment for plausibility. If the total time-investment exceeded 8 h per day, the reviewer could be discredited or downweighted.

A reviewer's total number of reviews (in a given year) and total time spent reviewing could also be used to limit a single person's influence.

### ***Reviewers can be evaluated by the predictive power of their reviews***

A reviewer who signs a review or rating links a little piece of his or her reputation to a paper. This is a gamble. Say the review was positive. If the paper stands the test of time, then the reviewer's reputation rises a little. If the paper becomes discredited, the reviewer's reputation falls. Since every scientist rates many papers, a single erroneous judgment will not have a large effect. A reviewer's performance on a given evaluation can be estimated as a function of the existing evaluations at the time of submission of the evaluation and the evaluations accumulated up to the present moment: Performance could be judged as high if the reviewer's judgment stands the test of time, and especially high if this evaluation was made early and/or diverged from existing evaluations when it was entered. This criterion can be formalized in an information-theoretic framework.

The OE system will enable scientists to make visible contributions by evaluating others' work. As a result, reviewing will be a competitive, public activity, that strongly impacts one's reputation as a scientist. Some scientists will contribute to the evaluation more than others. In fact, the system would enable some scientists to specialize in this particular form of meta-science. The system will fundamentally change the way science progresses: scientists will want to attach their reputations to the developments they truly believe in. Looking wisely ahead with deep intuition will be rewarded over following shallow trends.

### **WILL INSTANT PUBLISHING NOT DESTROY THE CONSTRUCTIVE PROCESS OF REVIEW AND REVISION?**

***No, revisions will still be possible in the proposed system, and they will often include improvements made in response to reviews***

A revision will take precedence over the original version of the paper in that it will be the version most visibly presented to readers. However, the entire history of the paper, including the original version, all revisions, and all evaluative meta-information will remain openly accessible and separately citable in perpetuity. The authors have no right or ability to remove this record.

If the authors decide to submit a revision of their paper, the revision will require re-review (as is the case in the current system for major revisions). The ratings and reviews of the original paper will not automatically transfer to the revision. If the revision is important to the field, it will be re-evaluated by enough scientists (likely including some of the original reviewers). If the revision is less important, it will not be as broadly and deeply evaluated as the original version, but can still serve to provide the most up-to-date version of the paper and address the reviews of the original.

The authors are free to refuse to revise their paper if other projects are of greater importance to them. When the authors disagree with reviews, they can publish responses to the reviews (as meta-reviews), which may contain further experimental results, along with ratings of the reviews. PEFs may utilize higher-order reviews in weighting the ratings of the first-order reviews.

Responses to reviews are simply reviews referring to other reviews, thus utilizing the same infrastructure as reviews of papers and meta-reviews contributed by other scientists. Author responses to reviews and will provide an important function complementary to that of a revision.

### ***Scientists will be highly motivated to seek informal feedback before publication of the original paper***

A paper, once published, can never be erased from the crystallized record of scientific history. Moreover, the attention the community grants to a new paper upon publication so as to evaluate it may not be reduplicated for a revision. This creates a strong motivation for scientists to publish only work they can stand by in the long run. Scientists will therefore seek informal constructive criticism before initial publication to a greater degree than currently. For example, in addition to presenting the project at a conference they may post the paper on a blog or share it with selected researchers by email a few weeks before publication. This informal round of review and revision will reduce the noise in the crystallized record.

### **CAN ALTERNATIVE METRICS, INCLUDING USAGE STATISTICS, SOCIAL-NETWORK INFORMATION, AND LINK-BASED IMPORTANCE INDICES, SERVE TO PRIORITIZE THE LITERATURE?**

Yes, alternative metrics derived from usage statistics, from links, and from the social web will play an increasing role in steering the attention of both the general public and the scientific community (Neylon and Wu, 2009; Priem and Hemminger, 2010; see also this collection: Birukou et al., 2011; Walther and van den Bosch, 2012; Zimmermann et al., 2012). However, evaluating science also requires conscious judgment by experts. In addition to the informal and fleeting buzz of the social web, we therefore need a system to collect and analyze explicit peer reviews and ratings.

Algorithms like PageRank (used by Google to prioritize search results) can provide overall importance indices, and can be modified to rely more heavily on some links (e.g., citations from scientific papers) than others. In usage and link-based importance indices, however, positive and negative attention adds to the visibility of the content. Explicit judgments, such as the "desired-impact" rating suggested above, provide a complementary signal that will be important in science. In contrast perhaps to other domains like art and entertainment, science will always rely on explicit peer judgment.

### **CAN RESEARCH BLOGGING SERVE THE FUNCTION OF OPEN PEER REVIEW, AND PERHAPS EVEN OF SCIENTIFIC PUBLISHING IN GENERAL?**

Research blogging fills an important gap: between informal discussions and formal publications (Harnad, 1990). Unlike a private informal discussion, a blog is publicly accessible. Unlike a scientific paper, a blog post can be altered or removed from public access. Blog posts are also often anonymous, whereas papers are signed and author-authenticated. These more fluid properties of blogs make for their unique contribution to scientific culture. However, the very fluidity of blogs also makes them inadequate as the sole vessel of scientific publishing. In particular, blogging lacks the quality of "crystallization" (Figure 9). A scientific publication needs to be crystallized in the sense that it is a constant

historical record that can be accessed permanently and therefore cited.

Blogs are science's short-term memory (**Figure 10**). They enable more intuitive and divergent reasoning. The crystallized literature is science's long-term memory, which enables more analytical and convergent reasoning. Crystallized scientific publications include papers and reviews. Reviews are crystallized publications that serve mainly to evaluate one or several other crystallized publications. Crystallized publications are digitally authenticated documents that reference other scientific publications.

The web's equivalent of a citation is a link. Links are versatile and fast, but there is no mechanism to ensure that they will continue to work in perpetuity. In fact, such a mechanism would rob the web of a key feature: plasticity. While the web world of blogs is fast and flexible, it is also fleeting and this is a good thing. As a complement to the web, however, we need a crystallized scientific record. Links here are citations of papers identified by digital object identifiers, which are guaranteed to be maintained in perpetuity. Links crossing the boundary between these two worlds are desirable. Scientific posts (i.e., a web document such as a blog post) will use web-links to other non-crystallized resources and in addition they will cite the crystallized record. Conversely, scientific papers (i.e., crystallized publications) will rely on citations to ground themselves in the crystallized scientific record and can additionally utilize web-links, with the understanding that these may become defunct.

#### WHAT IS THE ROLE OF PUBLISHING COMPANIES AND JOURNALS IN THE PROPOSED SYSTEM?

This proposal affirms the importance of the scientific paper and the process of peer review as essential elements of scientific publishing. The current function of the journal in administering peer review, selecting content, and providing access to related papers in context will be more fluidly served by web-portals that present a portion of the literature, prioritized by PEFs. The future system will be designed by scientists, independent of publishing companies. This reflects the fact that the key functions of access and

evaluation can be served at a higher level of quality and at lower costs than in the current system.

However, for-profit scientific publishers will have new opportunities to offer services that will legitimately contribute to science and society. The publication and review of specialized scientific papers might no longer depend on for-profit publishers, but their services can contribute to communicating the most important scientific findings beyond the confines of a highly specialized scientific audience. As an example of this challenge, let's consider the role currently played by the high-prestige publications *Nature* and *Science*.

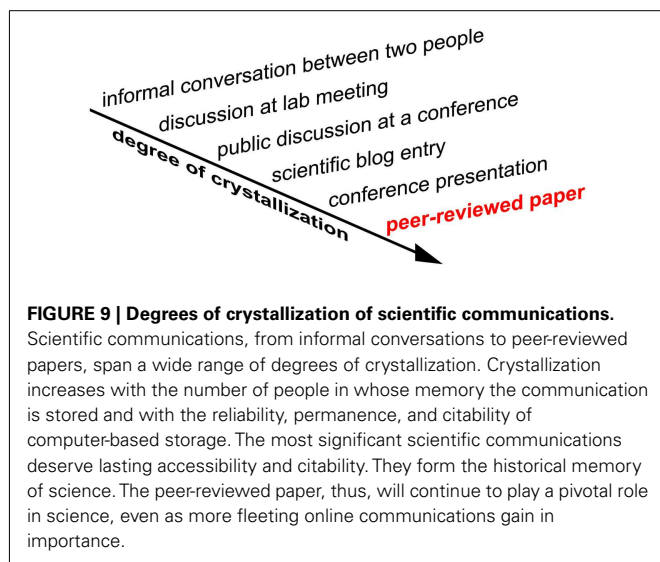
*Nature* and *Science* strive to reach a broad scientific audience with groundbreaking new science. They succeed in publishing many important papers. However, they use classical peer review, so their evaluation mechanism suffers from non-transparency (secret review) and from a lack of evaluative evidence (2–5 reviews). As a result, *Science* and *Nature*, despite having the highest standard in the industry, do not quite live up to their promise of publishing only groundbreaking work. Conversely, they miss out on groundbreaking work published elsewhere (because it was either not submitted to them or rejected by them). In addition, primary research papers in *Nature* and *Science* do not typically succeed at communicating their results to a broad audience.

In the future, a for-profit publisher could utilize the OE system, develop its own PEF for selecting content, and produce a high-prestige publication that fully succeeds (1) at presenting only groundbreaking science and (2) at communicating it to a broader audience. The content of such a general science magazine would not be primary reports of new scientific findings. Rather the publisher would select independently published studies that have turned out to be groundbreaking, relying on the broader, deeper, and more reliable evidence from OE. The original authors would then be invited to write a piece communicating the science more broadly (cf. the "Focused Review" format of *Frontiers*). Since the scientific validity and significance has already been established, the publisher's role would be to ensure readability and didactic quality of text and visuals. Copy editing and professional artwork and layout, as provided by publishing companies, are non-essential for primary research reports, but valuable for the broader communication of groundbreaking findings.

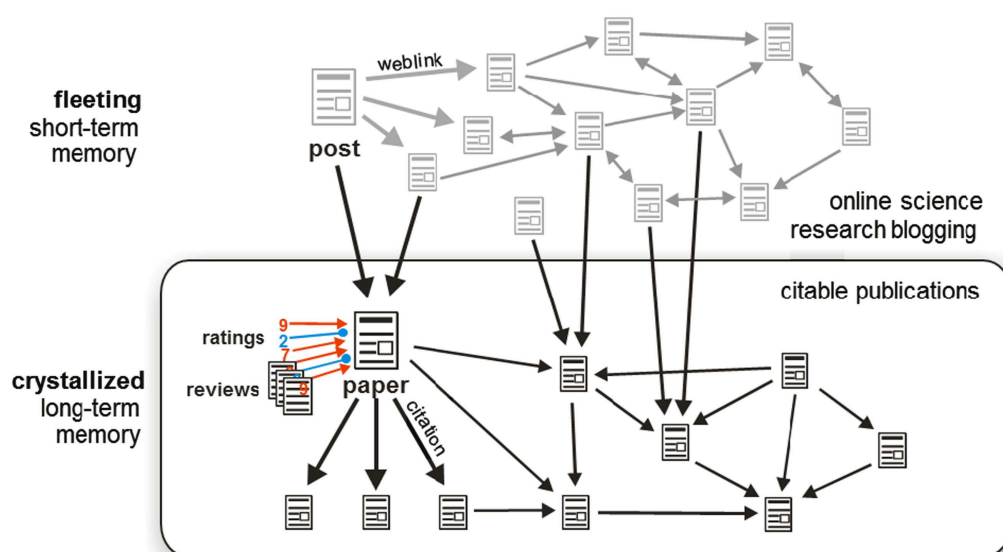
#### HOW CAN WE REALISTICALLY TRANSITION TO THE PROPOSED SYSTEM?

Transitioning to a radically different system is difficult. Clearing the slate and starting from scratch, i.e., *revolution*, is often politically and logistically unrealistic. In addition, no matter how brilliant and detailed our vision for the future system, it is bound to fall short of anticipating all complications encountered during implementation. Our vision might even be fundamentally flawed, in which case revolutionary change would be a catastrophic mistake.

Transitioning through *evolution*, on the other hand, is not always possible. The present system may be stuck in a local optimum, where any small changes worsen the situation. This could be among the reasons for the persistence of the current system of scientific publishing. Senior scientists and editors who appreciate the subtle checks and balances of the current system may feel







**FIGURE 10 | Fleeting online science and the crystallized scientific record.**

Much of online science is fleeting. For example, a link to a blog post becomes obsolete when the owner removes the post. This is as it should be. Research blogs serve as science's short-term memory. However, science also needs a long-term memory, a crystallized and permanently citable historical record. This function is served by the peer-reviewed literature. Note that fleeting

online science and the crystallized record interact intensely as bloggers refer to papers and blogs inspire new studies that later become part of the scientific record. However, while blogs link to other blogs (gray arrows) and cite papers (black downward arrows), scientific papers mainly cite other scientific papers (black arrows), because links to online science are less dependable in the long term.

that suggestions for change are naive and would not improve the situation.

Fortunately, there is a continuous path toward fundamental change of the scientific publishing system. To make change, we need to open up not only access, but also evaluation. Access and evaluation are the two major functions a publishing system must provide. With OA on the rise, evaluation, i.e., the stamp of approval implicit to acceptance of a paper in a journal of a given level of prestige, is the essential product the scientific publishers are selling today. Once scientists take on the challenge of envisioning, implementing, and using an independent and general OE system, change is underway.

An independently built OE system can evaluate the entire literature, including papers published under the current system, which appear in traditional journals. The tipping point is reached when the evaluations provided by the OE system are perceived as more reliable and authoritative than journal prestige as an indication of a new paper's quality. At this point, scientists will no longer be dependent on journals to publish their work.

The key challenge therefore is for the scientific community to converge on a vision for OE. This will require alternative proposals to be explored in detailed papers and to be widely discussed. We hope that the collection of visions presented in this Research Topic will contribute to this process.

#### WHO IS RESPONSIBLE FOR DESIGNING AND CONTINUALLY IMPROVING THE PUBLISHING SYSTEM?

It's up to scientists to design and continually improve the future publishing system. Providing access and evaluation of the

literature is properly construed as a key methodological challenge for science. Science tackles other difficult methodological challenges by means of methodological studies and a literature documenting the results. We also need a literature, both theoretical and empirical, exploring methods for OE.

So far scientists have largely left the design and justification of the evaluation process to journals and publishing companies. However, the evaluation system is a core component of science itself. It determines the confidence we can have in scientific findings. It steers the attention of the scientific community and affects public policy decisions. The evaluation system, therefore, must be designed by scientists. The behavioral, cognitive, computational, and brain sciences are best prepared to take on this task, which will involve social and psychological considerations, software design, and modeling of the network of scientific papers and their interrelationships. We need a literature that illuminates how we can bring science and statistics to the evaluation process.

The larger challenge is to design the collective cognition of the scientific community and its interaction with web-based artificial intelligence. OE is a core component of this collective cognitive system. Designing OE requires us to study (1) the individual scientist's motivation, cognition, and interaction with web-based human-computer interfaces, (2) the consequences of enabling different forms of individual influence on the system, (3) the dynamics of the entire system as a social network, (4) mechanisms for combining evaluations from many individual scientists so as to prioritize the literature, (5) the network of papers (nodes) and citations (links) and potential automatic inference methods (e.g., Bayesian belief propagation) that can be applied to this

network to assess the validity of the claims in the context of their interrelationships.

### SHOULDN'T WE STRIVE FOR AN EVEN MORE RADICAL VISION OF COLLABORATIVE SCIENCE?

Yes, we should. Web collaboration is bound to revolutionize the way science is done (Nielsen, 2009, 2011). Scientific teams collaborating on a problem will be distributed around the world and as the process becomes transparent throughout, traditional divisions will blur and might well evaporate. These divisions include

the temporal division between the stages of doing the science, of publishing it, and of review and reception; the division of communications between collaborative communication among the team and publication of the results; and the social division between team members (i.e., coauthors) and the audience of scientists exposed to the results. However, even when this dream has become a reality, we will still need a permanent record of scientific papers and explicit peer judgments. The present proposal focuses on this permanent record, but fits well into a larger vision of fluid, open, collaborative science on the web.

### REFERENCES

- Bachmann, T. (2011). Fair and open evaluation may call for temporarily hidden authorship, caution when counting the votes, and transparency of the full pre-publication procedure. *Front. Comput. Neurosci.* 5:61. doi:10.3389/fncom.2011.00061
- Birukou, A., Wakeling, J. R., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K., et al. (2011). Alternatives to peer review: novel approaches for research evaluation. *Front. Comput. Neurosci.* 5:56. doi:10.3389/fncom.2011.00056
- Boldt, A. (2010). *Extending ArXiv.org to Achieve Open Peer Review and Publishing*. Available at: <http://arxiv.org/abs/1011.6590>
- Brees, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. Technical Report MSR-TR-98-12. Microsoft Research. Microsoft Corporation.
- Florian, R. V. (2012). Aggregating post-publication peer reviews and ratings. *Front. Comput. Neurosci.* 6:31. doi:10.3389/fncom.2012.00031
- Frishauf, P. (2009). Reputation systems: a new vision for publishing and peer review. *J. Particip. Med.* 1, e13a.
- Godlee, F. (2002). Making reviewers visible openness, accountability, and credit. *JAMA* 287, 2762–2765.
- Goldberg, D. (1992). Using collaborative filtering to weave an information tapestry. *Commun. ACM* 35, 61–70.
- Greaves, S., Scott, J., Clarke, M., Miller, L., Hannay, T., Thomas, A., et al. (2006). Nature's trial of open peer review. *Nature*. Available at: <http://www.nature.com/nature/peer-review/debate/nature05535.html>
- Groves, T. (2010). Is open peer review the fairest system? Yes. *BMJ* 341, c6424.
- Harnad, S. (1990). Scholarly skywriting and the prepublication continuum of scientific inquiry. *Psychol. Sci.* 1, 342–343.
- Harnad, S. (1997). Learned inquiry and the net: the role of peer review, peer commentary and copyright. *Learned Publishing* 11, 283–292.
- Harnad, S. (2010). The open challenge: a brief history. *Public Serv. Rev. Eur. Sci. Technol.* 9, 13–15.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* 102, 16569–16572.
- Khan, K. (2010). Is open peer review the fairest system? No. *BMJ* 341, c6425.
- Kravitz, D. J., and Baker, C. I. (2011). Toward a new model of scientific publishing: discussion and a proposal. *Front. Comput. Neurosci.* 5:55. doi:10.3389/fncom.2011.00055
- Kreiman, G., and Maunsell, J. (2011). Nine criteria for a measure of scientific output. *Front. Comput. Neurosci.* 5:48. doi:10.3389/fncom.2011.00048
- Lee, C. (2012). Open peer review by a selected-papers network. *Front. Comput. Neurosci.* 6:1. doi:10.3389/fncom.2012.00001
- Neylon, C., and Wu, S. (2009). Article-level metrics and the evolution of scientific impact. *PLoS Biol.* 7, e1000242. doi:10.1371/journal.pbio.1000242
- Nielsen, M. (2009). Doing science in the open. *Phys. World* 22, 30–35.
- Nielsen, M. (2011). *Reinventing Discovery: The New Era of Networked Science*. Princeton: Princeton University Press, 280.
- Pöschl, U. (2010). Interactive open access publishing and peer review: the effectiveness and perspectives of transparency and self-regulation in scientific communication and evaluation. *LIBER Q.* 19, 293–314.
- Pöschl, U. (2012). Multi-stage open peer review: scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation. *Front. Comput. Neurosci.* 6:33. doi:10.3389/fncom.2012.00033
- Priem, J., and Hemminger, B. (2010). *Scientometrics 2.0: Toward New Metrics of Scholarly Impact on the Social Web*. Available at: <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874/2570> [First Monday, 15].
- Priem, J., and Hemminger, B. M. (2012). Decoupling the scholarly journal. *Front. Comput. Neurosci.* 6:19. doi:10.3389/fncom.2012.00019
- Pulverer, B. (2010). Transparency showcases strength of peer review. *Nature* 468, 29–31.
- Sandewall, E. (2012). Maintaining live discussion in two-stage open peer review. *Front. Comput. Neurosci.* 6:9. doi:10.3389/fncom.2012.00009
- Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). "Collaborative filtering recommender systems," in *International Journal of Electronic Business*, Vol. 2, eds P. Brusilovsky, A. Kobsa, and W. Nejdl (Berlin Heidelberg: Springer), 77.
- Smith, J. (1999). The deconstructed journal – a new model for academic publishing. *Learned Publishing* 12, 79–91.
- Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *J. R. Soc. Med.* 99, 178–182.
- Smith, R. (2009). In search of an optimal peer review system. *J. Particip. Med.* 1, e13.
- van Rooyen, S., Delamothe, T., and Evans, S. J. W. (2010). Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial. *Br. Med. J.* 341, c5729.
- van Rooyen, S., Godlee, F., Evans, S., Black, N., and Smith, R. (1999). Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomized trial. *Br. Med. J.* 318, 23–27.
- Walsh, E., Rooney, M., Appleby, L., and Wilkinson, G. (2000). Open peer review: a randomised controlled trial. *Br. J. Psychiatry* 176, 47–51.
- Walther, A., and van den Bosch, J. F. (2012). FOSE: a framework for open science evaluation. *Front. Comput. Neurosci.* 6:32. doi:10.3389/fncom.2012.00032
- Ware, M. (2011). Peer review: recent experience and future directions. *New Rev. Inf. Networking* 16, 23–53.
- Wichert, J. M., Kievit, R. A., Bakker, M., and Borsboom, D. (2012). Letting the daylight in: reviewing the reviewers and other ways to maximize transparency in science. *Front. Comput. Neurosci.* 6:20. doi:10.3389/fncom.2012.00020
- Zimmermann, J., Roebroek, A., Uludag, K., Sack, A. T., Formisano, E., Jansma, B., et al. (2012). Network-based statistics for a community driven transparent publication process. *Front. Comput. Neurosci.* 6:11. doi:10.3389/fncom.2012.00011

**Conflict of Interest Statement:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 May 2012; paper pending published: 22 June 2012; accepted: 18 September 2012; published online: 17 October 2012.

Citation: Kriegeskorte N (2012) Open evaluation: a vision for entirely transparent post-publication peer review and rating for science. *Front. Comput. Neurosci.* 6:79. doi: 10.3389/fncom.2012.00079

Copyright © 2012 Kriegeskorte. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and subject to any copyright notices concerning any third-party graphics etc.