frontiers in
COMPUTATIONAL NEUROSCIENCE

REVIEW ARTICLE
published: 03 April 2012
doi: 10.3389/fncom.2012.00020

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[1]. 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.

[1]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; Templer, 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

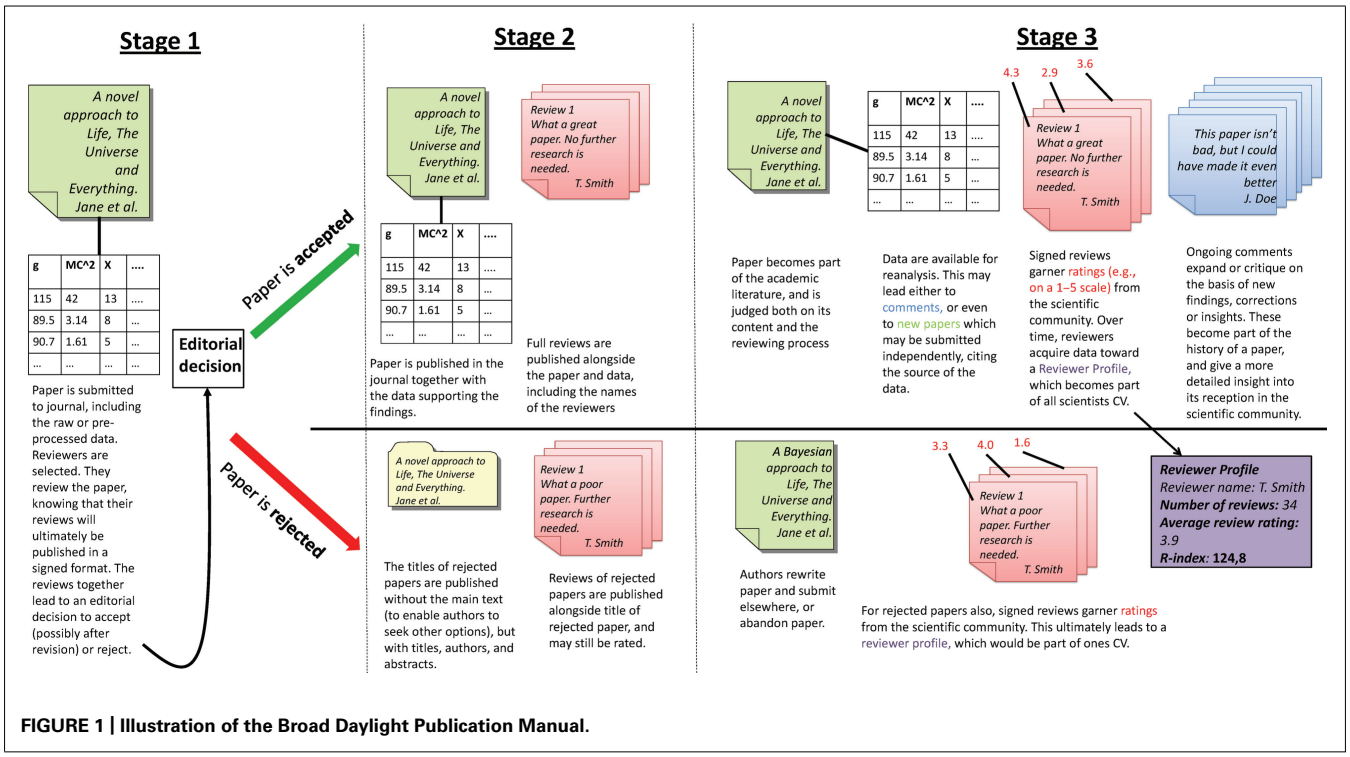

FIGURE 1 | Illustration of the Broad Daylight Publication Manual.

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[2], Neurosynth[3], and the BrainMap Project[4]. 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.

[2]http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml

[3]http://neurosynth.org

[4]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 Sherrat, 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 preprocessing of neuro-imaging data should be documented in ways that enable replication on the basis of the raw data, whereas preprocessed 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., Helantera, 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., Michalakis, 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., Gaggar, 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.

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 Sherrat, 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 nearsighted 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., Freckelton, 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., Winkielman, 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*