RESEARCH ARTICLE

Bias in peer review: a case study [version 1; peer review: 2 approved with reservations]

Richard Walker ¹, Beatriz Barros², Ricardo Conejo², Konrad Neumann³, Martin Telefont⁴

¹Frontiers, EPFL Innovation Park, Lausanne, 1015, Switzerland

²Department of Computer science, University of Malaga, Malaga, 29071, Spain

³Institute of Medical Biometrics and Clinical Epidemiology, Charité, Universitätsmedizin, Berlin, 10117, Germany

⁴Blue Brain Project, EPFL ENT CBS BBP / HBP, Geneva, 1211, Switzerland

V1 First published: 22 Jan 2015, 4:21
<https://doi.org/10.12688/f1000research.6012.1>

Latest published: 10 Jun 2015, 4:21
<https://doi.org/10.12688/f1000research.6012.2>

Abstract

Peer review is the "gold standard" for evaluating journal and conference papers, research proposals, on-going projects and university departments. However, it is widely believed that current systems are expensive, conservative and prone to various forms of bias. One form of bias identified in the literature is "social bias" linked to the personal attributes of authors and reviewers. To quantify the importance of this form of bias in modern peer review, we analyze three datasets providing information on the attributes of authors and reviewers and review outcomes: one from Frontiers - an open access publishing house with a novel interactive review process, and two from Spanish and international computer science conferences, which use traditional peer review. We use a random intercept model in which review outcome is the dependent variable, author and reviewer attributes are the independent variables and bias is defined by the interaction between author and reviewer attributes. We find no evidence of bias in terms of gender, or the language or prestige of author and reviewer institutions in any of the three datasets, but some weak evidence of regional bias in all three. Reviewer gender and the language and prestige of reviewer institutions appear to have little effect on review outcomes, but author gender, and the characteristics of author institutions have large effects. The methodology used cannot determine whether these are due to objective differences in scientific merit or entrenched biases shared by all reviewers.

Keywords

Peer review, bias, nationality, gender, language, prestige, random intercept model, authors, reviewers

Open Peer Review

Approval Status

	1	2
version 2		
(revision)
10 Jun 2015
version 1		
22 Jan 2015

1. Lutz Bornmann, Administrative

Headquarters of the Max Planck Society,
Munich, Germany

2. Jigisha Patel, Biomed Central Ltd, London, UK

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **Research on
Research, Policy & Culture** gateway.

Corresponding author: Richard Walker (richard.walker@epfl.ch)

Competing interests: Frontiers was a partner in the SISOB project and contributed to the research described in this paper. The first author (RW) is a part-time employee of Frontiers.

Grant information: The research described in this paper was supported by a grant from the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement 7PM- 266588 (SISOB).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2015 Walker R *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Walker R, Barros B, Conejo R *et al.* **Bias in peer review: a case study [version 1; peer review: 2 approved with reservations]** F1000Research 2015, 4:21 <https://doi.org/10.12688/f1000research.6012.1>

First published: 22 Jan 2015, 4:21 <https://doi.org/10.12688/f1000research.6012.1>

Introduction

Peer review is the “gold standard” for the evaluation of journal and conference papers, research proposals, on-going projects and university departments and there is a strong consensus in the scientific community that it improves the quality of scientific publications^{1,2}. As reported by Armstrong, “journal peer review is commonly believed to reduce the number of errors in published work, to serve readers as a signal of quality and to provide a fair way to allocate journal space”³. Surveys of authors and expert reviewers^{4–6} show that this view is widely held. However, many members of the scientific community also believe that peer review is expensive, conservative and prone to bias^{2,7–18}. Critics point to the major delays it introduces into the publication process^{17,19}, to biases against particular categories of papers (e.g. studies challenging conventional wisdom²⁰; replication studies^{21,22} and studies reporting negative results^{12,23}), to the unreliability of the review process^{23–25}), to its inability to detect errors and fraud²⁶, and to unethical practices by editors and reviewers^{27,28}.

Another common criticism of peer review is that it is prone to “social bias”²⁹: certain categories of reviewer may have conscious or unconscious biases for or against particular categories of author (e.g. authors of a particular gender, authors coming from institutions in a particular geographical area, authors from institutions in English-speaking or non-English speaking countries, authors belonging to low or high prestige institutions). Attempts to measure these biases have given contrasting results. For example, an experimental study by Lloyd³⁰ shows that manuscripts with female author names have a far higher acceptance rate when they are reviewed by female rather than male reviewers (62% vs. 21%). Along the same lines, a widely quoted study of grant awards in Sweden³¹ suggests that proposals from male candidates receive systematically higher evaluations than those from female candidates with similar academic records, a result confirmed by a recent follow-up study¹⁵. Similarly, a study of the introduction of double blind review in the journal *Behavioral Ecology* reports that blinding reviewers to author gender led by an increase in papers with female first authors, absent in other very similar journals³². However, the interpretation of this study has been contested³³, and some *ex post* analyses of publication patterns indicate that differences in the way male and female reviewers review papers do not affect review outcomes³⁴. Other studies have found robust effects of author gender on peer review results but do not determine whether these are the result of bias or of differences in the scientific merit of articles submitted by male and female authors³⁵.

Other forms of social bias (e.g. bias for or against authors from particular geographical areas, language bias, bias in favor of authors from high prestige institutions) have been studied less frequently than gender bias but have also produced contrasting results, comprehensively reviewed by Lee and colleagues²⁹. Several studies have shown, for instance, that journals favor authors located in the same country as the publisher^{36,37}. Thus, a retrospective analysis by Link³⁸ suggests that American reviewers are significantly more likely to accept a paper by another American author than a paper by an author of a different nationality. However, other studies indicate that American reviewers are actually more lenient to non-American than to American authors³⁹. A study by Tregenza and colleagues⁴⁰ shows that papers by authors from institutions in countries

whose native language is English are more successful than papers by authors from institutions in other countries, an opinion shared by many commentators (for example⁴¹). However, the authors explicitly state that this is not necessarily a sign of bias and other studies have not found the same effect (e.g.⁴²). Peters and Ceci⁴³ report a quasi-experiment demonstrating that papers by authors from high-prestige institutions have a significantly higher chance of acceptance than similar papers by authors with less prestigious affiliations and participants in surveys of authors are reported to believe in such an effect⁴⁴. However, to the knowledge of the authors, there is little observational evidence for this effect.

Attempts to remedy the weaknesses of traditional peer review have led to a diversification of peer review practices, for instance through the use of author-blind and non-selective review, the removal of traditional reviewer anonymity, and the introduction of various forms of community review. To date, however, there have been few attempts to measure their effectiveness. Furthermore, many past studies of bias in peer review are relatively old and it is not clear whether or how far biases detected in the past have been affected by changes in social attitudes. In response to these concerns, we analyze data for the peer review systems used in Frontiers (an open access publishing house which uses a novel interactive review process), three computer science conferences (CAEPIA2003, JITEL2007, and SINTICE07) held in Spain between 2003 and 2007 and four international computer science conferences (AH2002, AIED2003, ICALP2002 and UMAP2011), held between 2002 and 2011.

Frontiers

Frontiers is a large open access scientific publisher, which published its first journal in 2007. In January 2015, Frontiers had a portfolio of 49 open-access journals with over 50,000 researchers serving on its editorial boards and more than 380 academic specialty sections. Papers are published within these sections. Each paper is assigned to a scientist acting as the editor who coordinates the review process, and is responsible for publication/rejection decisions. Reviewers are selected automatically, based on the match between their individual specialties, and key words in articles submitted for review (or can also be assigned manually by the editor). During the review process, reviewers remain anonymous. However, accepted papers carry the names of their reviewers. This gives reviewers a strong incentive not to accept papers before they have reached a good level of quality.

The Frontier’s review system is designed not so much to *select* papers as to *improve* the quality of papers in a collaborative, interactive dialog between authors and reviewers. At the beginning of the review process, reviewers are asked to answer a series of open questions concerning different aspects of the paper (see [Table 1](#)). The precise set of questions depends on the nature of the paper (original research, review paper, etc.). Authors answer reviewer questions through an interactive forum. This possibility reduces misunderstandings and can significantly accelerate the submission process.

In addition to replying to open questions, reviewers can express their overall evaluation of the paper on a range of numerical scales (see [Figure 1](#)). However, the use of these scales is not mandatory. Nowhere in the process do papers receive an aggregate numerical score. The final decision to accept or reject a paper is taken by the

Table 1. A sample of questions to reviewers used in the Frontiers review process.

1) Are there any objective errors in the results? If yes, please specify.
2) Is the work ethical in your opinion?
3) Was the research carried out in accordance with established animal use practices?
4) Is any use of human subjects performed according to acceptable standards?
5) Has the clinical trial been registered in a public trials registry?
6) Should an accession number of nucleotide/amino acid sequences be included?
7) Should an accession number for microarray data be included?
8) Does the article describe experiments using a select agent or toxin?

Figure 1. Summary scales used in the Frontiers review system.

journal editor, based on the overall results of the interactive review process. Acceptance rates are high. Of the papers in the Frontiers database that had reached a final publication/rejection decision on the date when we extracted the data, 91.5% were published and only 8.5% were rejected.

IEEE (Spain) and IEEE (international)

The conferences in these two datasets all used WebConf (<http://WebConf.iaia.lcc.uma.es>), a computerized system for managing the submission and review of conference papers. WebConf was developed by a team from Malaga University led by one of the authors (RC). WebConf implements a classical review process similar to the processes used by Springer, Elsevier and other large commercial publishers. Conference contributions are usually reviewed by three independent reviewers, occasionally by two or four. Reviewers are chosen by the conference program chair who draws on a database of potential experts called the program committee. In general, the program committee is made up of authors who have previously submitted papers in a particular area of research and have expressed their willingness to act as reviewers. The WebConf system suggests a list of potential reviewers based on the degree of matching between paper topics and reviewers field of expertise. The final selection is based on the judgment of the program chair.

Reviewers express their opinion of a paper in a conference-specific review form in which they assign scores to the paper on a number of

separate scales, covering key areas of evaluation (typically including soundness, originality, clarity etc.) and textual comments. Scores on individual scales are usually expressed in terms of categories (typically: poor, fair, good, excellent). The final publication decision depends on the program chair. If one of the reviewers expresses a strongly negative view of a paper, it will typically be rejected. In cases where there is a very significant difference in reviewers' opinions, the program chair can ask for an additional review. Acceptance rates vary between a minimum of 29.3% and a maximum of 87%. The combined acceptance rate for the seven conferences in the WebConf database was 57.9%.

Methods
Data

Frontiers. The Frontiers database includes details of all authors and reviewers for all scientific papers submitted to Frontiers (N=8,565) between June 25, 2007 and March 19, 2012, the name of the journal to which the paper was submitted, the article type (review, original research etc.), the name and institutional affiliations of the authors and reviewers of specific papers, individual reviewer scores for the summary scales shown in Figure 1, and the overall review result (accepted/rejected). At the time of the analysis, 2,926 papers had not completed the review process and were excluded. In another 1,089 cases, reviewers had not assigned numerical scores to the paper, which could not therefore be considered. Our final analysis used 9,618 reviews, for 4,549 papers. Most of the papers in the database come from the life sciences. The majority of authors and reviewers

come from Western Europe and Northern America. However, the database contains a substantial number of authors and reviewers from other parts of the world.

Spanish computer science conferences (IEEE-Spain). This dataset includes details of 1,131 reviews referring to 411 papers submitted to three IEEE conferences (CAEPIA2003, SINTICE2007 and JTEL2007). The majority of authors and reviewers for these papers come from institutions in Spain and Portugal. The data provided include the name of the conference to which the contribution was submitted, the type of contribution (poster, short paper, full paper etc.) the name, gender and institutional affiliations of the authors and reviewers of specific contributions, individual reviewers scores and the final decision (accepted/rejected). All the papers in the database are in the area of computer science.

International computer science conferences (IEEE-International). This dataset provides data for 2,194 reviews, referring to 793 papers submitted to four IEEE conferences (AH2002, AIED2003, ICALP2002 and UMAP2011), managed using WebConf and involving authors and reviewers from all over the world. This dataset provided the same data collected for IEEE-Spain.

Normalization of author and reviewer names

Names of authors and reviewers were canonized: accent and symbols were removed; double spaces replaced by single spaces, and upper-case characters replaced with lower-case characters. Names were rewritten in the normalized form <first name, last name>. Intermediate names were omitted.

Normalization of institution names

Names of institutions (universities, research institutions, companies) were canonized as above. After normalization, the name of the institution was recoded using the first three words in the full name.

Gender assignment

Neither the Frontiers nor the WebConf databases included data for author and reviewer gender. In the Frontiers case, gender was inferred semi-automatically in a multistep process. First we matched the first names contained in our database to an open source dictionary providing genders for more than 40,000 first names used in different countries (gender-1.0.0.tgz, downloadable from <http://pecl.php.net/package/gender>). We then used volunteers of different nationalities (Chinese, Egyptian, Indian, Japanese, Korean, Turkish) to assign genders to first names not contained in the dictionary. Additional names were assigned by manually searching for specific authors and reviewers on Google and Facebook. At the end of this process, we were able to assign genders to 9,472 out of 11,729 authors and reviewers. The majority of unassigned names were Asian (mainly Chinese). This is a potential source of bias in our analysis. Genders for WebConf authors and reviewers were inferred using a similar procedure.

Assignment of countries/geographical area/English vs. non-English speaking world

To ascertain possible country and regional biases, we assigned each author and reviewer to the country of the institution to which they were affiliated. Authors with multiple affiliations were assigned to

the country of the first affiliation listed. The USA, UK, Scotland, N. Ireland, Ireland, Australia, Canada, and New Zealand were classified as English-speaking countries. All other countries were classed as non-English speaking.

University rankings

Authors' and reviewers' affiliated institutions were classified in terms of their position in the 2012 Shanghai academic ranking of world universities for the life sciences (<http://www.shanghairanking.com/FieldLIFE2012.html>) (Frontiers) and for computer sciences (<http://www.shanghairanking.com/SubjectCS2012.html>) (WEB-CONF). Rank was coded as a categorical variable with values 1 (rank 1–50), 2 (rank 51–200), 3: unclassified. Given the process used to normalize institution names, it is possible that a small number of institutions were misclassified.

Calculation of review scores

Frontiers. The Frontiers review process produces a very low rejection rate. This means that the database used for our study contained relatively few rejected papers (N=478). To create a more informative indicator of reviewers' evaluations, we computed for each paper the average of the scores expressed by each individual reviewer for the summary scales shown in Figure 1. A comparison between the distributions of scores for rejected and published papers (see Figure 2) clearly demonstrates the validity of the indicator. However, it should be noted that the indicator is a construction of the authors and played no role in the review process.

WebConf. The WebConf system asks each reviewer to assign an overall score to the paper he/she has just reviewed. Scores are expressed on a scale of 0 to 10.

Statistical analysis

For the purposes of the study, we define bias as the interaction terms δ_{ij} in the random intercept model:

$$y = b + \mu_{ij} + \beta_{ij} A_i + \gamma_{ij} R_j + \delta_{ij} A_i R_j + \varepsilon \quad (1)$$

where y denotes the score given in the review, b denotes the random intercept, i indexes properties of authors, j indexes properties of reviewers, and ε is the error term. This method is similar but not identical to the method proposed in⁴⁵.

Given a factor F , such as *region*, the variables A_i and R_j are indicator (dummy) variables indicating that the first author and reviewer belong to categories i and j of factor F , respectively, that is:

$$\text{IF one author belongs to category } i \text{ of } F, A_i = 1, \text{ ELSE } A_i = 0 \quad (2)$$

AND

$$\text{IF one reviewer belongs to category } j \text{ of } F, R_j = 1, \text{ ELSE } R_j = 0 \quad (3)$$

Thus β_{ij} is the fixed effect of author category, γ_{ij} is the fixed effect of reviewer category and δ_{ij} is the fixed effect of the interaction between author and reviewer category. Since the expected value of the random intercept b is 0, the fixed effects allow us to estimate the following mean scores:

Figure 2. Distribution of reviewer scores for published and rejected papers in the Frontiers dataset.

$G_{ij} = \mu_{ij} + \beta_{ij} + \gamma_j + \partial_{ij}$ author in category i , reviewer in category j

$G_{ij} = \mu_{ij} + \beta_{ij}$ author in category i , reviewer *not* in category j

$G_{ij} = \mu_{ij} + \gamma_j$ author category i , reviewer in category j

$G_{ij} = \mu_{ij}$ author *not* in category i , reviewer *not* in category j .

We define bias B_{ij} of reviewers from category j of factor F towards authors from category i by the expression:

$$B_{ij} = (G_{ij} - G_{ij}) - (G_{ij} - G_{ij}) = \partial_{ij} \quad (4)$$

Since the intercept and main effects cancel out, bias is the interaction term ∂_{ij} and does not depend on the main effects β_{ij} and γ_j . In other words, it is independent of any general tendency of authors in category i to write better papers than other authors, or of any tendency of reviewers in category j to give generally higher scores. In this setting, reviewers from category j are *biased in favor of* authors from category i , if $B_{ij} > 0$ and are *biased against* authors from category i if $B_{ij} < 0$. Bias is significant at a level α if we can reject the null hypothesis:

$$H_0 : B_{ij} = 0$$

Otherwise we assume absence of bias.

The majority of papers in our databases had multiple authors. In preliminary studies, we explored statistical models that used this data in different ways: (i) the model used only the properties of the

first author, (ii) the model used only the properties of the last author, (iii) the model considered the properties of all the authors. The three approaches yielded similar results (data not shown). In what follows, we apply the first method, unless otherwise stated.

Example

To illustrate the concept of bias defined in (4), consider a factor with two levels such as gender. Let i and j denote female (F), and γ and γ male (M). Then the terms $G_{ij} = G_{FF}$, $G_{ij} = G_{FM}$, $G_{ij} = G_{MF}$, and $G_{ij} = G_{MM}$, have the following meanings:

G_{FF} : mean score when first author and reviewer are female

G_{FM} : mean score when first author is female and reviewer is male

G_{MF} : mean score when first author is male and reviewer is female

G_{MM} : mean score when first author and reviewer are male

If we assumed that papers by female authors have the same quality as papers by male authors,

$$G_{FF} - G_{FM} > 0 \quad (5)$$

would imply that female reviewers are biased in favor of female authors and

$$G_{MF} - G_{MM} > 0 \quad (6)$$

would imply that female reviewers are biased in favor of male authors.

However, we cannot make this assumption. We therefore conclude that female reviewers are biased for or against female authors ($B_{FF} \neq 0$) only if (5) and (6), yield different results. If both were equal and positive, we conclude that females give higher scores than men regardless of the gender of the author. If female reviewers have a positive bias, this always implies a negative bias on the part of male reviewers and vice versa. By construction, this method cannot detect biases shared by all reviewers (e.g. a bias against female authors, or authors from a particular geographical region, shared by all reviewers, regardless of gender or geographical origin).

It should also be noted that, in our example, if $B = (G_{FF} - G_{FM}) - (G_{MF} - G_{MM}) > 0$ the following statements are equivalent:

- Female reviewers are biased in favor of female authors
- Female reviewers are biased against male authors
- Male reviewers are biased against female authors
- Male reviewers are biased in favor of male authors

Similarly, if $B < 0$ we can draw the following equivalent conclusions:

- Female reviewers are biased against female authors
- Female reviewers are biased in favor of male authors
- Male reviewers are biased in favor of female authors
- Male reviewers are biased against male authors

Results

Gender

The automatic gender assignment program assigned genders to first authors and reviewers for 8,114 reviews from Frontiers, 1,131 from IEEE (Spain) and 2,194 from IEEE (International). The relative proportions of male first authors and reviewers (male authors: 70.0%–73.9%; male reviewers: 75.2–79.3%) were similar in all three datasets. In the Frontiers and IEEE (International) datasets, means scores for male first authors were significantly higher than those for female first authors (Frontiers: difference=0.07, $p=0.034$; IEEE International: difference=0.28, $p=0.001$). The IEEE (Spain) dataset showed the same pattern but the difference was not significant (difference=0.39, $p=0.40$). Reviewer gender had no significant effect on review scores in any of the datasets. In none of the data sets did the interaction between author and reviewer gender have a significant effect. In brief, none of the datasets showed evidence of gender bias. The significance of these results will be examined in the discussion. Complete data for the analysis can be found in **Data Files 1–4**.

Region

The analysis examined the role of the region of first author and reviewer institutions in determining review scores and tested for possible regional bias. Authors and reviewers were grouped into 11 geographical regions (Africa, Australia/New Zealand, Central America/Caribbean, Central Asia, Eastern Asia, Eastern Europe,

Middle East/North Africa, South America, Southern Asia, Southern Europe, and Western Europe) according to the location of their respective institutions. To avoid problems with the convergence of the mixed model algorithm and to guarantee the statistical power of the analysis, pairs of first author/reviewer regions with less than 30 reviews were discarded. Distributions of author and reviewer regions differed significantly among the three datasets. In the Frontiers and International IEEE datasets, the majority of authors and reviewers came from institutions in North America and Western Europe, while the majority of authors and reviewers in the Spanish IEEE dataset, came from institutions in Southern Europe.

All three datasets showed large differences in the scores of first authors from different regions (see **Table 2**), which were statistically significant even after correction for multiple hypothesis testing. In the Frontiers dataset, authors from N. America scored significantly higher than authors from all other regions, whereas authors from E. Asia, E. Europe, S. Asia and Southern Europe scored significantly lower. In the IEEE (Spain) dataset, authors from Southern Europe scored higher than authors from other regions, whereas authors from North America scored lower. In the IEEE (international dataset) authors from N. America again scored significantly higher than authors from other regions, while authors from Africa and Central Asia scored lower.

In the Frontiers and IEEE (Spain) datasets, we detected no significant differences in scores from reviewers from different regions (not shown). In the IEEE (International) dataset reviewers from Australia/NZ and from Southern Asia gave scores that were significantly higher than the average for reviewers from other regions and reviewers from Western Europe gave scores that were significantly lower (not shown).

Table 2. Differences in mean scores for first authors from institutions in different geographical regions with $p \leq 0.050$.

Frontiers		Bonferroni corrected alpha=0.005		
	N	Mean score region	Mean score other regions	p-value
Eastern Asia	501	7.20	7.40	0.009
Eastern Europe	82	6.85	7.40	0.003
North America	3,733	7.50	7.32	<0.001
Southern Asia	86	6.71	7.40	<0.001
Southern Europe	705	7.21	7.41	0.003
IEEE Spain		Bonferroni corrected alpha=0.0125		
Central America	49	5.57	6.11	0.048
South America	36	5.37	6.11	0.017
Southern Europe	976	6.20	5.39	<0.001
IEEE (International)		Bonferroni corrected alpha=0.0028		
Eastern Asia	161	5.22	5.67	0.039
North America	503	6.11	5.50	<0.001
Southern Europe	441	5.29	5.72	0.004

To test for bias, we applied the random intercept model to all author/reviewer region pairs with more than 30 reviews (see Table 3). After application of the Bonferroni correction for multiple hypothesis testing, the Frontiers dataset showed no evidence of interaction between author and reviewer region and the other two datasets showed only very limited evidence (IEEE – Spain: bias of reviewers from S. Europe in favor of authors from E. Asia; IEEE – International: bias of North American reviewers in favor of authors from Eastern Asia). In none of the datasets did we find any evidence for regional biases previously reported in the literature (e.g. bias of North American reviewers in favor of North American authors). We conclude that regional bias has only a limited effect on review scores and that those biases that do exist are idiosyncratic to particular review systems. Full data for the analysis can be found in **Data Files 5–8**.

Language

We hypothesized that reviewers could be biased against papers written by authors who were not native English speakers. We, therefore, analyzed potential reviewer bias for and against papers, whose first authors came or did not come from institutions in English-speaking countries. As a further test, we analyzed potential bias for and against papers, which had, or did not have, at least one author belonging to an institution in an English speaking country.

The Frontiers and the IEEE (International) datasets, both included large numbers of authors and reviewers, from institutions in native English-speaking and from non-English-speaking countries. In contrast, approximately 97% of the authors and reviewers in the IEEE

(Spain) dataset came from Spain. Since none of the papers with an English-language first author, and only one paper with at least one English author, were reviewed by an English-language reviewer, it was not possible to measure bias using the random intercept model. This dataset was therefore discarded from the subsequent analysis.

In the remaining datasets, papers with a first author from an institution in a non-English-speaking country scored significantly lower than papers with first authors from institutions in an English speaking country, regardless of whether they were reviewed by native English-speaking or non-native English speaking reviewers (Frontiers: difference=-0.22, $p<0.001$; IEEE International: difference=-0.54, $p<0.001$). Reviewer language had no significant effect on score (Frontiers: difference=0.04, $p=0.040$; IEEE International: difference=0.01; $p=0.80$). In neither case did we find a significant interaction between author and reviewer language (not shown). Results for papers with at least one author from an institution in an English-speaking country were similar. Details of the analysis are shown in **Data Files 9–11**.

Ranking of author and reviewer institutions

Reviewers from institutions with high academic prestige could be biased in favor of authors from other high prestige institutions and against authors from lower ranking institutions. To test this possibility, we classified all authors and reviewers in the three datasets by the position of their institutions in the Shanghai classifications, as described earlier.

The Frontiers and the IEEE International datasets both contained significant numbers of authors and reviewers from universities in all three categories. However, nearly all the authors and reviewers in the IEEE Spain dataset came from universities in category 3. Given the lack of data for authors and reviewers from higher-ranking institutions, this dataset was excluded from the subsequent analysis.

In the Frontiers dataset, authors from universities in category 1 scored significantly higher than authors from category 2 (difference=0.16, $p=0.016$) and from category 3 (difference=0.21, $p<0.001$), regardless of the origin of the reviewer. No significant difference was observed between the scores of authors in category 2 and 3 or between scores given by reviewers from institutions in different categories. In the IEEE International dataset, authors from universities in categories 1 and 2 both scored significantly higher than authors in category 3 (category 1: difference in scores=0.97, $p<0.001$; category 2: difference in scores=0.89, $p<0.001$) but there was no significant difference in their own respective scores (not shown). There was no significant difference between the scores given by reviewers in category 1 and the scores given by reviewers in categories 2 and 3. However, reviewers in category 2 gave significantly higher scores than reviewers in category 3 (difference in scores=0.28, $p<0.001$). Neither dataset showed a significant interaction between the prestige of author and reviewer institutions. In the IEEE (International) dataset, reviewers from institutions in category 1 gave higher scores to authors in category 3 than to authors in category 1, but the difference was not statistically significant (difference=0.48, $p=0.059$). Full details of the analysis are given in **Data Files 12–14**.

Table 3. Bias in mean scores for specific author-reviewer region pairs with uncorrected p-values ≤ 0.05 ; * shows significant bias after application of the Bonferroni correction for multiple hypothesis testing (Frontiers $\alpha = 0.001$; IEEE-Spain $\alpha < 0.0125$; IEEE-International $\alpha < 0.0028$).

Frontiers				
Author	Reviewer	N	Difference	p-Value
E. Asia	E. Asia	130	0.35	0.047
E. Asia	S. Europe	34	-0.58	0.028
N. America	E. Europe	36	-0.63	0.034
IEEE Spain				
Author	Reviewer	N	Bias	p-Value
E. Asia	S. Europe	46	2.34	0.005*
IEEE International				
Author	Reviewer	N	Bias	p-Value
E. Asia	N. America	31	1.06	<0.001*
E. Asia	W. Europe	60	-0.46	0.048
N. America	W. Europe	223	-0.29	0.04
S. Europe	S. Europe	109	0.4	0.023
W. Europe	S. Europe	153	-0.3	0.05
W. Europe	W. Europe	344	0.28	0.026

Discussion

The results of the study (see [Table 4](#)) show that the scores received by papers in peer review depend strongly on the characteristics of the first author (gender, geographical location, language and prestige of the author's institution). In summary, male authors receive higher scores than female authors, authors from some geographical regions receive higher scores than authors from others; authors from institutions in English-speaking countries receive higher scores than authors in non-English-speaking countries; authors from high prestige institutions receive higher scores than authors from lower-prestige institutions. In contrast, we find no evidence that scores are affected by the personal characteristics of reviewers, no significant interactions between author and reviewer gender, language, and institutional prestige and only limited evidence of an interaction between author and reviewer region. In brief, the study provides little evidence for bias, at least in the sense in which bias is defined in our study (see below).

The review systems considered in the study are very different. The majority of papers in the Frontiers dataset came from the life sciences; all the papers in the IEEE datasets were from specialized areas of computer science. Frontiers adopts a novel interactive review process; the conferences in the IEEE (Spain) and the IEEE (International) datasets used a traditional approach. Authors and reviewers in the Frontiers and the IEEE International datasets come

from all over world. The IEEE (Spain) dataset is dominated by authors and reviewers from Southern Europe. Despite these differences, analysis of the three datasets gave very similar results. This suggests that the findings of this study could be valid for a broad range of peer review systems. The large size of the datasets used in the analysis (in total 12,943 reviews of 5,743 papers) provides additional evidence of robustness. The main differences between the datasets were in their patterns of regional bias, which are different in each dataset. Unfortunately the many differences between the Frontiers and the IEEE systems make it impossible to untangle the roles of different contributory factors.

The finding that author characteristics have a significant effect on review scores is compatible with two distinct explanatory hypotheses. The first is that papers submitted by authors with a particular characteristic (e.g. authors from institutions in a particular region) are, on average, of higher scientific merit than papers by authors with different characteristics (e.g. authors from institutions in other regions). The second is that reviewers share a general bias against authors with particular characteristics, regardless of their own characteristics (e.g. reviewers from institutions in English and non-English speaking countries share a bias against authors from non-English speaking countries). The methodology of the study cannot distinguish between these hypotheses; in fact, the only way to detect generalized bias would be through experimental

Table 4. Summary of main findings.

Finding	Frontiers	IEEE (Spain)	IEEE (International)
Male first authors achieve higher scores than female first authors	X		X
There is no significant difference between scores from male and female reviewers	X	X	X
There is no evidence of gender bias (significant interaction between author and reviewer gender)	X	X	X
In a small number of cases, first authors from particular geographical regions score significantly higher or significantly lower than authors from all other regions.	X	X	X
There are no significant differences between scores given by reviewers from different regions	X	X	X
There is little or no regional bias (little or no evidence for interaction between author and review region)	X	X	X
Authors from institutions in English-speaking countries score higher than authors in non-English speaking countries	X	N/A	X
There is no significance difference between scores from reviewers from institutions in English and non-English speaking countries	X	N/A	X
There is no significant language bias (no interaction between language of author and reviewer institutions)	X	N/A	X
Scores for authors from institutions with high Shanghai rankings are significantly higher than scores for authors from lower ranking institutions	X	N/A	X
There is no significant difference between scores from reviewers from institutions in different Shanghai categories	X	N/A	X
There is no bias in terms of the Shanghai category of author and reviewer institutions (no significant interaction between them)	X	N/A	X

studies, comparing review scores when reviewers are blinded to particular characteristics of the author to the scores given when they are not (for example^{30,43}). Such studies are extremely valuable. However, their experimental nature means that they are unable to demonstrate the existence or absence of bias in operational review systems. This suggests that observational and experimental methods are complementary, and that it will not be possible to gain a complete picture of bias in peer review without using both methods.

The study found no evidence for social bias in terms of author gender, and the language and prestige of author institutions and only weak evidence of regional bias. The findings for gender match results from a previous study which used the interaction between author and reviewer characteristics as a measure of bias⁴⁶. However, it apparently contradicts results from previous studies showing significant bias, with respect to gender (e.g.^{30,43}) as well as for region (e.g.³⁸), and institutional prestige (e.g.⁴³). Given that the majority of studies showing bias are relatively old, it is possible that changes in social attitudes have reduced or eliminated some of the biases they observed. In some cases, (e.g.³¹) studies measured effects that were independent of reviewer characteristics, which, as explained earlier, are invisible to the methodology used in this study. In this case, the results of our study would complement rather than contradict previous results. We suggest, nonetheless, that, at least in the case of gender, it is implausible that modern female reviewers are biased against female authors. The most parsimonious explanation of our results is that, with the possible exception of regional biases, social bias plays at most a minor role in determining review outcomes.

Our study also found no significant differences between the scoring patterns of different categories of reviewer. These results, which were valid for all three datasets, contrast with previous findings showing significant differences in scoring practices between male and female reviewers^{34,47} and between reviewers from different countries⁴⁷. However, there have been relatively few studies on this issue, and even these do not show a major impact of reviewer characteristics on the final outcomes of the review process³⁴. Taken together, these results suggest that editors should not be over-concerned with the gender, language or institutional affiliation of the reviewers they choose for particular papers, though it could be useful to ensure a good regional balance among reviewers.

Our study does not evaluate the full set of potential biases described in the peer review literature. For instance, we do not consider confirmation bias or alleged reviewer biases in favor of positive results, sophisticated experimental and statistical methodology, or against interdisciplinary studies, replication studies, etc. These are important

limitations. A second limitation is that the study methodology has so far been tested with just three peer review systems, all applying to scientific papers. It is possible that other forms of peer review, such as peer review of grant applications, are subject to different forms of bias.

In conclusion, our study shows the important role of authors' personal characteristics in determining the scores received in peer review, but finds little evidence for bias due to interactions between author and reviewer characteristics. These findings do not rule out generalized bias against authors with specific characteristics or forms of bias not considered in the study.

Data availability

To protect the identities of authors and reviewers, the source data used for this study have not been made public.

Author contributions

RW, RC, BB and KN conceived the study; RW wrote the paper with contributions from KN and RC; MT prepared the Frontiers data for use in the analysis and implemented *ad hoc* software for this purpose; BB and RC prepared the Spanish and the International Computer Science Conference datasets; KN designed and implemented the statistical analysis. All authors contributed critical comments and agreed to the final content of the article.

Competing interests

Frontiers was a partner in the SISOB project and contributed to the research described in this paper. The first author (RW) is a part-time employee of Frontiers.

Grant information

The research described in this paper was supported by a grant from the European Union's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement 7PM- 266588 (SISOB).

I confirm that the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors gratefully acknowledge the technical support and constructive criticism received from members of the Frontiers team, in particular Fred Fenter, Kamila Markram, and Costanza Zucca. We further acknowledge the contributions of participants in SISOB meetings, where we presented preliminary versions of this study, in particular those from Ulrich Hoppe and Sandor Soos. A preliminary version of this paper was incorporated in "SISOB Deliverable 9.3: Study of Enhanced Evaluation".

References

1. Nature: **Pros and cons of open peer review**. *Nature Neuroscience: Nat Neurosci*. 1999; 2(3): 197–198.
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Ware M: **Peer review: benefits, perceptions and alternatives**. *Publishing Research Consortium*. 2008.
[Reference Source](#)
3. Armstrong JS: **Management Science: What Does It Have to do With Management or Science?** *Mark Bull*. 1998; 9: 1–13.
[Reference Source](#)
4. Bradley JV: **Pernicious publication practices**. *Bull Psychon Soc*. 1981; 18(1): 31–34.
[Publisher Full Text](#)

5. Fletcher RH, Fletcher SW: **Evidence for the effectiveness of peer review.** *Sci Eng Ethics.* 1997; **3**(1): 35–43.
[Publisher Full Text](#)
6. MacNealy MS, Speck BW, Clements N: **Publishing in technical communication journals from the successful author's point of view.** *Tech Commun.* 1994; **41**(2): 240–259.
[Reference Source](#)
7. Nature: **Can peer review police fraud?** *Nat Neurosci.* 2006; **9**(2): 149.
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Jefferson T, Rudin M, Brodney Folse S, *et al.*: **Editorial peer review for improving the quality of reports of biomedical studies.** *Cochrane Database Syst Rev.* 2007; (2): MR000016.
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Bohannon J: **Who's Afraid of Peer Review?** *Science.* 2013; **342**(6154): 60–65.
[PubMed Abstract](#) | [Publisher Full Text](#)
10. Bornmann L: **Scientific peer review.** *Annu Rev Inf Sci Technol.* 2011; **45**(1): 197–245.
[Publisher Full Text](#)
11. Ellison G: **Is peer review in decline?** *Econ Inq.* 2011; **49**(3): 635–657.
[Publisher Full Text](#)
12. Emerson GB, Warne WJ, Wolf FM, *et al.*: **Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial.** *Arch Intern Med.* 2010; **170**(21): 1934–9.
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Smith R: **Classical peer review: an empty gun.** *Breast Cancer Res.* 2010; **12**(Suppl 4): S13.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Suls J, Martin R: **The air we breathe: A critical look at practices and alternatives in the peer-review process.** *Perspect Psychol Sci.* 2009; **4**(1): 40–50.
[Publisher Full Text](#)
15. Sandström U, Hällsten M: **Persistent nepotism in peer-review.** *Scientometrics.* 2008; **74**(2): 175–189.
[Publisher Full Text](#)
16. Alberts B, Hanson B, Kelner KL: **Reviewing peer review.** *Science.* 2008; **321**(5885): 15.
[PubMed Abstract](#) | [Publisher Full Text](#)
17. Benos DJ, Bashari E, Chaves JM, *et al.*: **The ups and downs of peer review.** *Adv Physiol Educ.* 2007; **31**(2): 145–152.
[PubMed Abstract](#) | [Publisher Full Text](#)
18. Smith R: **Peer review: a flawed process at the heart of science and journals.** *J R Soc Med.* 2006; **99**(4): 178–182.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Armstrong JS: **Peer Review for Journals: Evidence on Quality Control, Fairness, and Innovation.** *Sci Eng Ethics.* 1997; **3**(1): 63–84.
[Publisher Full Text](#)
20. Campanario JM: **Rejecting and resisting Nobel class discoveries: accounts by Nobel Laureates.** *Scientometrics.* 2009; **81**(2): 549–565.
[Publisher Full Text](#)
21. Kerr S, Tolliver J, Petree D: **Manuscript Characteristics Which Influence Acceptance for Management and Social Science Journals.** *Acad Manage J.* 1977; **20**(1): 132–141.
[Publisher Full Text](#)
22. Campanario JM: **Peer review for journals as it stands today-Part 1.** *Sci Commun.* 1998; **19**(3): 181–211.
[Publisher Full Text](#)
23. Mahoney MJ: **Publication prejudices: An experimental study of confirmatory bias in the peer review system.** *Cognit Ther Res.* 1977; **1**(2): 161–175.
[Publisher Full Text](#)
24. Kravitz RL, Franks P, Feldman MD, *et al.*: **Editorial peer reviewers' recommendations at a general medical journal: are they reliable and do editors care?** *PLoS One.* 2010; **5**(4): e10072.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Herron DM: **Is expert peer review obsolete? A model suggests that post-publication reader review may exceed the accuracy of traditional peer review.** *Surg Endosc.* 2012; **26**(8): 2275–2280.
[PubMed Abstract](#) | [Publisher Full Text](#)
26. Schroter S, Black N, Evans S, *et al.*: **Effects of training on quality of peer review: randomised controlled trial.** *BMJ.* 2004; **328**(7441): 673.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Travis GDL, Collins HM: **New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System.** *Sci Technol Hum Values.* 1991; **16**(3): 322–341.
[Publisher Full Text](#)
28. Sieber JE: **Improving the peer-review process relies on understanding its context and culture.** *Nature.* 2012.
[Publisher Full Text](#)
29. Lee CJ, Sugimoto CR, Zhang G, *et al.*: **Bias in peer review.** *J Am Soc Inform Sci Technol.* 2013; **64**(1): 2–17.
[Publisher Full Text](#)
30. Lloyd ME: **Gender factors in reviewer recommendations for manuscript publication.** *J Appl Behav Anal.* 1990; **23**(4): 539–543.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Wenneker C, Wold A: **Nepotism and sexism in peer-review.** *Nature.* 1997; **387**(6631): 341–343.
[PubMed Abstract](#) | [Publisher Full Text](#)
32. Budden AE, Tregenza T, Aarssen LW, *et al.*: **Double-blind review favours increased representation of female authors.** *Trends Ecol Evol.* 2008; **23**(1): 4–6.
[PubMed Abstract](#) | [Publisher Full Text](#)
33. Webb TJ, O'Hara B, Freckleton RP: **Does double-blind review benefit female authors?** *Trends Ecol Evol.* 2008; **23**(7): 351–353.
[PubMed Abstract](#) | [Publisher Full Text](#)
34. Gilbert JR, Williams ES, Lundberg GD: **Is there gender bias in JAMA's peer review process?** *JAMA.* 1994; **272**(2): 139–142.
[PubMed Abstract](#) | [Publisher Full Text](#)
35. Bornmann L, Mutz R, Daniel HD: **Gender differences in grant peer review: A meta-analysis.** *J Informetrics.* 2007; **1**(3): 226–238.
[Publisher Full Text](#)
36. Daniel HD, Russey WE: **Guardians of science: Fairness and reliability of peer review: VCH Weinheim.** 1993.
[Publisher Full Text](#)
37. Ernst E, Kienbacher T: **Chauvinism.** *Nature.* 1991; **352**: 560.
[Publisher Full Text](#)
38. Link AM: **US and non-US submissions: an analysis of reviewer bias.** *JAMA.* 1998; **280**(3): 246–247.
[PubMed Abstract](#) | [Publisher Full Text](#)
39. Marsh HW, Jayasinghe UW, Bond NW: **Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability.** *Am Psychol.* 2008; **63**(3): 160–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
40. Tregenza T: **Gender bias in the refereeing process?** *Trends Ecol Evol.* 2002; **17**(8): 349–350.
[Publisher Full Text](#)
41. Herrera AJ: **Language bias discredits the peer-review system.** *Nature.* 1999; **397**(6719): 467.
[PubMed Abstract](#) | [Publisher Full Text](#)
42. Loonen MP, Hage JJ, Kon M: **Who benefits from peer review? An analysis of the outcome of 100 requests for review by Plastic and Reconstructive Surgery.** *Plast Reconstr Surg.* 2005; **116**(5): 1461–1472.
[PubMed Abstract](#) | [Publisher Full Text](#)
43. Peters DP, Ceci SJ: **Peer-review practices of psychology journals: The fate of published articles, submitted again.** *Behav Brain Sci.* 1982; **5**(2): 187–195.
[Publisher Full Text](#)
44. Gillespie GW, Chubin DE, Kurzon GM: **Experience with NIH peer review: researchers' cynicism and desire for change.** *Sci Technol Hum Val.* 1985; **10**(3): 44–54.
[Publisher Full Text](#)
45. Bornmann L, Mutz R, Daniel HD: **How to detect indications of potential sources of bias in peer review: A generalized latent variable modeling approach exemplified by a gender study.** *J Informetrics.* 2008; **2**(4): 280–287.
[Publisher Full Text](#)
46. Jayasinghe UW, Marsh HW, Bond N: **A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings.** *J R Stat Soc Ser A Stat Soc.* 2003; **166**(3): 279–300.
[Publisher Full Text](#)
47. Borsuk RM, Aarssen LW, Budden AE, *et al.*: **To name or not to name: The effect of changing author gender on peer review.** *BioScience.* 2009; **59**(11): 985–989.
[Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 13 February 2015

<https://doi.org/10.5256/f1000research.6434.r7428>

© 2015 Patel J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jigisha Patel

Biomed Central Ltd, London, UK

In this paper the authors analyses the relationship between the attributes of authors and reviewers and reviewer outcomes in three datasets, one an innovative peer-review model and two which use traditional peer review. They find no evidence of gender or institutional bias and limited interaction between author and reviewer region.

Although social bias in peer review has been investigated in the past, I think this study is timely and the question of whether this previously found bias still exists with changing social norms and values is interesting. However, there is a need for some clarification of the methods before the significance of the authors' findings can be determined.

Major revisions

1. The authors are very clear that their study investigates 'bias' as defined by the interaction between author *and* reviewer attributes, or bias as a function of reviewer characteristics (ref 29). They make the point that their methodology cannot determine biases that are shared by all reviewers regardless of reviewer characteristics. In the Introduction, the authors switch between describing previous research on reviewer characteristics and previous research on author characteristics. For example, paragraph 2 is predominantly about studies of reviewer characteristics, but ends on author characteristics. Paragraph 3 appears to begins on author characteristics, but then cites studies on reviewer characteristics. It would be much easier for the reader to understand what contribution this study makes to the literature if the authors made a clearer distinction between current evidence on *reviewer* characteristics (including those studies which looked at both reviewer and author characteristics) and other research on bias, which would include research focused only on author characteristics.
2. As part of the peer review process for Frontiers reviewers can, if they wish, complete the ratings questionnaire shown in Fig 1 and it was these ratings that were used in the study analysis.

I think it would be useful for the authors to clarify the following:

- Do Frontiers reviewers complete the rating independently of each other before the collaborative process? If this is the case, it is not clear to me why individual scores for each paper were averaged for the Frontiers dataset, but apparently not for the other two datasets. If the aim of this study is to investigate the interaction between reviewer attributes and those of authors, wouldn't averaging the reviewer scores in this way confound this aim?
- Alternatively, if the ratings form is completed by reviewers after the collaborative process, how have the authors accounted for the potential confounding effect of the collaboration?
- Also, if reviewers complete the rating independently of each other, the characteristics of peer review for the Frontiers dataset used in this study are the same as that for the other datasets, i.e. it is single blind peer review. In all three datasets the reviewers are made aware of the authors names, but the authors do not know the reviewers' – is that correct?
- In the discussion the authors state that the findings of this study could be valid for a broad range of peer review systems. However, this study did not include the interactive component of Frontiers peer review process, or if it did, it is not clear how. All three datasets appear to have used the single blind system of peer review. This statement in the discussion should be rephrased.

Minor

Reference 31 is for a commentary. Can authors should provide the reference for the original Swedish study?

Can the authors provide data on the error rate for their gender assignment process?

I think the authors could provide a more informative title, for example, 'Bias in peer review: the interaction between reviewer and author characteristics.'

Please note, I do not have the expertise to comment on the model and statistical analysis used in this study.

Competing Interests: I am an employee of BioMed Central, an open access publisher.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 27 May 2015

Richard Walker

We have now submitted a new version of our paper, which we have revised in the light of your comments, which we found extremely useful and constructive, though they caused us significant extra work. We are particularly grateful for your suggestions concerning the organization of the introduction to the paper, which we have attempted to take on board, and for your points on quality control (which were in the same spirit as comments from Lutz

Bornmann. Thanks to these suggestions, we found a number of problems with the data, which had previously passed unnoticed.

- *Comment: The authors are very clear that their study investigates 'bias' as defined by the interaction between author and reviewer attributes, or bias as a function of reviewer characteristics (ref 29). They make the point that their methodology cannot determine biases that are shared by all reviewers regardless of reviewer characteristics. In the Introduction, the authors switch between describing previous research on reviewer characteristics and previous research on author characteristics. For example, paragraph 2 is predominantly about studies of reviewer characteristics, but ends on author characteristics. Paragraph 3 appears to begin on author characteristics, but then cites studies on reviewer characteristics. It would be much easier for the reader to understand what contribution this study makes to the literature if the authors made a clearer distinction between current evidence on reviewer characteristics (including those studies which looked at both reviewer and author characteristics) and other research on bias, which would include research focused only on author characteristics.*

This was an extremely useful suggestion. We have now reorganized our introduction to talk first about effects regarding author characteristics, then to interactions between author and reviewer characteristics and finally to the characteristics of reviewers. We believe the paper gains significantly in clarity from this reorganization.

- *As part of the peer review process for Frontiers reviewers can, if they wish, complete the ratings questionnaire shown in Fig 1 and it was these ratings that were used in the study analysis. Do Frontiers reviewers complete the rating independently of each other before the collaborative process? If this is the case, it is not clear to me why individual scores for each paper were averaged for the Frontiers dataset, but apparently not for the other two datasets. If the aim of this study is to investigate the interaction between reviewer attributes and those of authors, wouldn't averaging the reviewer scores in this way confound this aim?*

The ratings questionnaire is filled in in the initial non-interactive part of the review process. We have clarified this in the text.

- *Also, if reviewers complete the rating independently of each other, the characteristics of peer review for the Frontiers dataset used in this study are the same as that for the other datasets, i.e. it is single blind peer review. In all three datasets the reviewers are made aware of the authors names, but the authors do not know the reviewers' – is that correct?*

Yes this is correct. We have clarified our description of the Frontiers process to show that it is single-blind.

- *In the discussion the authors state that the findings of this study could be valid for a broad range of peer review systems. However, this study did not include the interactive component of Frontiers peer review process, or if it did, it is not clear how. All three datasets appear to have used the single blind system of peer review. This statement in the discussion should be rephrased*

The differences between the datasets concern not just the way the review is organized (interactive vs. non-interactive) but also the disciplines covered, and the geographical distribution of authors and reviewers. In the text, we clarify that all three review systems in the study are single blind. However, we hypothesize that reviewers preparing for an interactive process may behave differently from reviewers in a traditional review process.

- *Reference 31 is for a commentary. Can authors should provide the reference for the original Swedish study?*

As you correctly note, the reference (now reference 30) was published in the form of a commentary. De facto, however, the article represents the first publication of results from an original study, which, to our knowledge was not published elsewhere, prior to the date on which the commentary appeared.

- *Can the authors provide data on the error rate for their gender assignment process?*

On the basis of random sampling, we estimate errors rates of 7.5%, 0.0% and 5.2% for the Frontiers, IEEE (Spain) and IEEE (International) datasets respectively. We give these figures and discuss their significance in the text.

- *I think the authors could provide a more informative title, for example, 'Bias in peer review: the interaction between reviewer and author characteristics.'*

We agree with this suggestion and have revised our title accordingly. It now reads, "Personal attributes of authors and reviewers, social bias and the outcomes of peer review: a case study".

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 February 2015

<https://doi.org/10.5256/f1000research.6434.r7425>

© 2015 Bornmann L. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Lutz Bornmann

Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Munich, Germany

Peer review is the most important instrument for assessing scientific research. However, the instrument is not without drawbacks. As the most important weaknesses, a missing reliability,

fairness and predictive validity have been seen¹. The study of Walker, Barros, Conejo, Neumann and Telefont (2015) deals with the fairness of the peer review process: They investigated “social biases” in the processes of *Frontiers* - an open access publishing house with a novel interactive peer review process, and two peer review processes from Spanish and international computer science conferences.

The study is very interesting. I recommend that the authors revise the manuscript according to the following points:

1. On page 5, Walker, *et al.* describe the process of normalizing authors’ and reviewers’ names. Here, Walker, *et al.* should ensure that the names are completely cleaned: the same author and the same institutional unit should receive the same name. A general problem of this kind of data from peer review processes is that they are not cleaned and contain several name variants.
2. Walker, *et al.* present their results with the reporting of statistical significance information. I recommend that not only the statistical, but also the practical significance of the results (effect sizes) should be reported.^{2,3,4}
3. In the Discussion section the following major limitation of the study should be mentioned: The quality of the papers was not controlled. For example, Bornmann and Daniel (2009)⁵ and Bornmann and Daniel (2010)⁶ investigated the peer review processes of the *Angewandte Chemie – International Edition* and *Atmospheric Chemistry and Physics*. They considered citations for the single papers as a proxy for quality. Although citations measure only one part of quality (namely impact), it is more favorable to consider them than doing not. When examining the association of bias variables and peer review outcomes without controlling quality it is impossible to establish unambiguously whether a particular group of papers receives more favorable recommendations or decisions due to these variables, or if the more favorable recommendations and decisions are simply a consequence of the papers’ scientific quality.

References

1. Bornmann L: Scientific peer review. *Annual Review of Information Science and Technology*. 2011; **45** (1): 197-245 [Publisher Full Text](#)
2. Cohen J: Statistical power analysis for the behavioral sciences (2nd Ed).1988; **Lawrence Erlbaum Associates Hillsdale, NJ, USA**. [Reference Source](#)
3. Williams R: Using the margins command to estimate and interpret adjusted predictions and marginal effects.*The Stata Journal*. 2012; **12** (2): 308-311 [Reference Source](#)
4. Williams R, Bornmann L: The substantive and practical significance of citation impact differences between institutions: guidelines for the analysis of percentiles using effect sizes and confidence intervals. In Y. Ding, R. Rousseau & D. Wolfram (eds.), *Measuring scholarly impact: methods and practice*. 2014; **Springer, Heidelberg, Germany**: 258-281 [Reference Source](#)
5. Bornmann L, Daniel H-D: Reviewer and editor biases in journal peer review: an investigation of manuscript refereeing at Angewandte Chemie International Edition. *Research Evaluation*. 2009; **18** (4): 262-272 [Publisher Full Text](#)
6. Bornmann L, Daniel H: Do Author-Suggested Reviewers Rate Submissions More Favorably than Editor-Suggested Reviewers? A Study on Atmospheric Chemistry and Physics. *Plos One*. 2010; **5**

(10). [PubMed Abstract](#) | [Free Full Text](#) | [Publisher Full Text](#)

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 11 Feb 2015

Richard Walker

I think these comments are very relevant and we will take them into account in the next version of our paper. As concerns the specific points you raise:

1. In our study, authors' and reviewers' names are relevant for gender assignment. Institution names are critical for assignment of language, region, and institutional prestige to author and reviewer institutions. Informal checks on our cleaning process suggest that it does not introduce substantial errors into our analysis. In the next version of our paper we will introduce more formal checks.
2. We agree with the reviewer on this point. The next version of our paper will include estimates of the practical significance of our results.
3. We completely agree with the reviewer that some differences in scores may be due to differences in scientific quality and tried to make this point in our text. In the next version, we will attempt to clarify this issue.

Competing Interests: No competing interests were disclosed.

Author Response 27 May 2015

Richard Walker

Thank you for your comments, which we found extremely useful and constructive. Particularly useful was your suggestion that we should include effect sizes in our analysis. This suggestion has been implemented in this revised version of our paper. We would also like to thank you for your suggestions regarding quality control, complementing similar suggestions from Jigisha Patel. Thanks to these suggestions, we found a number of problems with the data, which had previously passed unnoticed. All these changes have been incorporated in the new version of the paper which we have just submitted.

- *On page 5, Walker, et al. describe the process of normalizing authors' and reviewers' names. Here, Walker, et al. should ensure that the names are completely cleaned: the same author and the same institutional unit should receive the same name. A general problem of this kind of data from peer review processes is that they are not cleaned and contain several name variants.*

Thank you this comment, which, together with comments in a similar vein from

Jigisha Patel led us to conduct a thorough review of our data. The review found that errors in the normalization of author, reviewer and institution names and missing values led to down-stream errors in automated gender assignment and in the assignment of university rankings. In the case of the university rankings we were able to correct a number of errors. In the case of the gender assignment, this would have been too onerous to be possible. Instead, as suggested by Jigisha Patel, we used sampling to check the error rates in our data, which we report in our text, together with a discussion of their significance.

- *Walker, et al. present their results with the reporting of statistical significance information. I recommend that not only the statistical, but also the practical significance of the results (effect sizes) should be reported.*^{2,3,4}

Thank you for this suggestion, which we have implemented in the revised version of our manuscript. We began by calculating Cohen's d for all our data. In the results and the discussion sections, we analyse the practical significance of the effects observed both in terms of the number of reviews concerned (which was always small) and the effects on publication decisions. In the cases we study, this was not large. However, in more selective review systems it could be larger.

- *In the Discussion section the following major limitation of the study should be mentioned: The quality of the papers was not controlled. For example, Bornmann and Daniel (2009)⁵ and Bornmann and Daniel (2010)⁶ investigated the peer review processes of the Angewandte Chemie – International Edition and Atmospheric Chemistry and Physics. They considered citations for the single papers as a proxy for quality. Although citations measure only one part of quality (namely impact), it is more favourable to consider them than doing not. When examining the association of bias variables and peer review outcomes without controlling quality it is impossible to establish unambiguously whether a particular group of papers receives more favourable recommendations or decisions due to these variables, or if the more favourable recommendations and decisions are simply a consequence of the papers' scientific quality*

We now make explicit reference to this limitation in our text. We go on to explain that most of the papers in our datasets have very few citations. This means, that in our case, we are unable to use citations as a proxy for quality.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research