Editorial

The significance of the peer review process against the background of bias: priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias

Tobias Opthof[a,*], Ruben Coronel[b,c], Michiel J. Janse[d]

[a]*Department of Medical Physiology, University Medical Center, PO Box 85060, 3508 AB Utrecht, The Netherlands*
[b]*Department of Cardiology, University Medical Center, Utrecht, The Netherlands*
[c]*Experimental and Molecular Cardiology Group, Academic Medical Center, Amsterdam, The Netherlands*
[d]*Editorial Office Cardiovascular Research, Academic Medical Center, Amsterdam, The Netherlands*

1. Introduction

Editors are facing larger numbers of submitted manuscripts than they can publish [1]. In their selection process of papers they depend on the advice of one or more peer reviewers [1]. It is thus important that the process is fair and as unbiased as possible.

The origin of 'peer review' dates back to 1752, when the Royal Society of London obtained the fiscal responsibility for the *Philosophical Transactions* [2]. There are few historical accounts of the evolution of editorial peer review [3]. Today, specialized research on the peer review system is only just starting to emerge and has been the topic of four world congresses [4–7]. This type of research focuses—amongst many other issues—on themes as whether or not masking the identity of authors to reviewers influences the reports of reviewers and whether or not anonymity of reviewers is a relevant topic [4–7].

Previous reports on the editorial process of *Radiology* [8] and the *Journal of Clinical Investigation* [9], have indicated that reviewers set markedly different standards in their appreciation of manuscripts. The concordance between reviewers on identical manuscripts is limited [9]. The fact that in the social sciences published papers have an almost 90% chance to be rejected when resubmitted to other journals casts doubt on the validity of the peer review system [10]. Such data are to our knowledge not available for the biomedical sciences. However, there is circumstantial evidence that peer review can successfully discriminate between manuscripts that have a greater chance to be cited in the future. Thus, Wilson showed that papers rejected by the *Journal of Clinical Investigation* were cited at lower frequency if published by other journals [11]. Manuscripts rejected by *Cardiovascular Research* were also cited at significantly lower frequency, even if published by journals with a higher impact factor [12].

Although there are clear indications that peer review can help selecting papers with a high scientific quality (if citation of papers is accepted as a parameter for scientific quality), this does not exclude bias in the process. Gender bias has been demonstrated in the peer review process of the *Journal of the American Medical Association*, although it was said not to have influenced the acceptance rates for male and female corresponding authors [13]. Peer review of grant proposals, however, appeared substantially biased by gender at the disadvantage of females [14].

Link [15] studied all original submissions to *Gastroenterology* in 1995 and 1996. It appeared that reviewers from the USA had a preference for manuscripts from the USA compared to reviewers from outside the USA. Such a difference was not seen for manuscripts outside the USA. Because the credibility of the peer review system seems pivotal for the (scientific) society we analyzed the material submitted to *Cardiovascular Research* between 1 October 1997 and 1 January 2002. More specifically, we analyzed (i) the predictive power of reviewer's priority recommendations for future citation, (ii) the predictive power of editor's ratings for future citation, (iii) concordance be-

*Corresponding author. Tel.: +31-30-253-8900; fax: +31-30-253-9036.
E-mail address:* t.opthof@med.uu.nl (T. Opthof).

tween reviewer's priority recommendations and editor's ratings and (iv) the occurrence of geographical bias.

We demonstrate that both (i) reviewer's recommendations and (ii) editor's ratings are positively correlated with citation, (iii) high editor's ratings in combination with high reviewer's priority recommendations are the strongest predictor for frequent citation. Finally, we demonstrate that geographical bias plays a role in the peer review process.

2. Can editors predict the priority assigned to manuscripts by reviewers ?

Soon after taking office on 1 June 1995 the number of submissions grew at a rate incompatible with the administrative power of our staff. We sought therefore for a sort of 'in house reject'. Because the urge to do so was based on practical considerations, we thought it fair to ask ourselves the question whether or not we could predict the priority assigned to manuscripts by a total of—in general—three reviewers. The answer was yes and no. 'Yes', because the overall relation between editor's rating and reviewer's priority was highly significant, but 'no' because the scatter was so substantial that it had little meaning for individual manuscripts. Between 1 October 1997 and 1 January 2002 we rated 3444 original manuscripts and correlated these with the averaged priority recommendations of the reviewers. Fig. 1 shows that the correlation was $Y=25.1+0.27X$ with a correlation coefficient of 0.236, which fulfills $P<0.0005$. The reviewers and editorial team assigned a score >50 to 24.5%, respectively, 27.3% of the manuscripts. Based on chance one would expect 6.7% of the manuscripts in the upper right quadrant of Fig. 1. In practice only 9.3% of the manuscripts were in this quadrant. The fact that 15.3% of manuscripts were in the upper left quadrant of Fig. 1 (low editor's score combined with high reviewer's score) made us to conclude that our predictive power was not impressive enough to reject manuscripts without sending them out to reviewers.

3. Can reviewers predict citation ?

Fig. 2 shows the relationship between the reviewer's priority score and citations obtained over a full period of 3
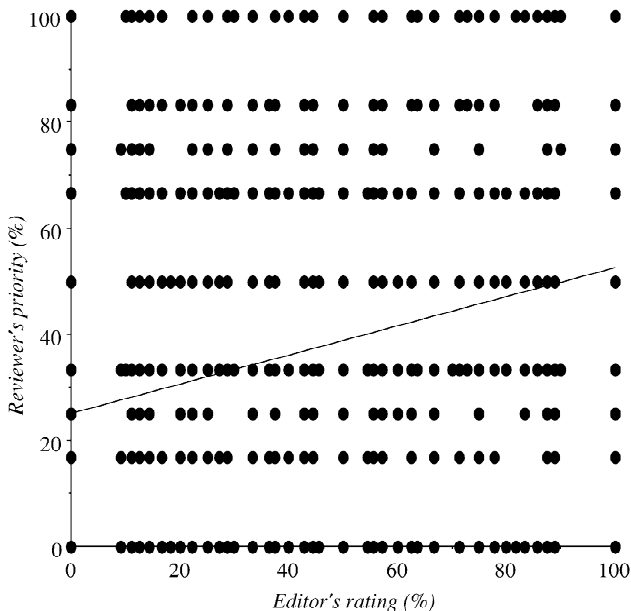
Fig. 1. Correlation between editor's ratings and reviewer's priority recommendations on 3444 original submissions between 1 October 1997 and 1 January 2002. Eight to 11 editors ranked manuscripts as pass or reject based on the abstract. This led to an editor's rating score between 0 and 100%. Three (78%) or two (22%) reviewer's priority recommendations which could either be high or low (or intermediate if the form was 'forced') together made the reviewer's priority score, which also varied between 0 and 100%. $Y=25.1+0.27X$, $r=0.236$, $P<0.0005$. See text for further details.
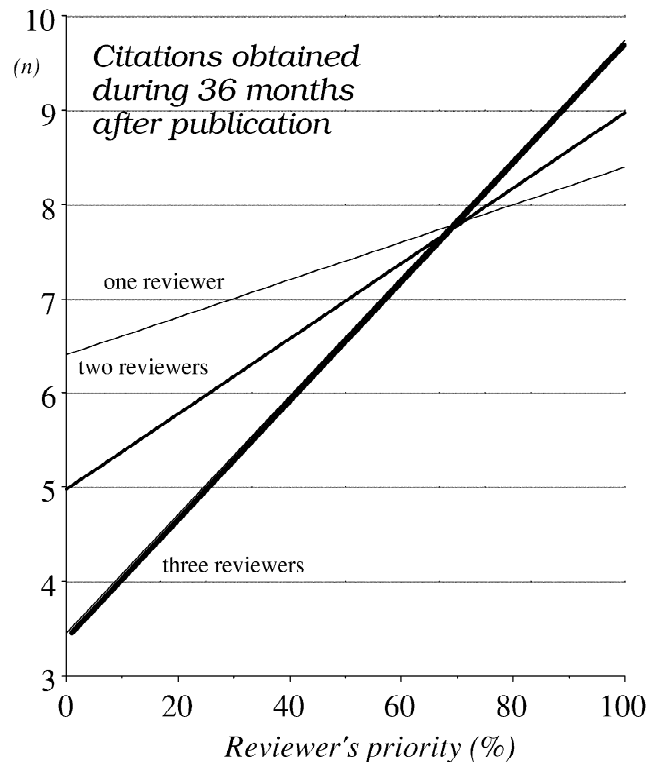
Fig. 2. Relationship between the priority recommendation of one, two and three reviewers on future citation over a period, which was exactly 3 years for each individual paper from a series of 183 papers published between April 1998 and September 1999 and cited between April 1998 and March 2001 (for the papers published in April 1998) and between September 1999 and August 2002 (for the papers published in September 1999). One reviewer: $Y=6.6+0.02X$, $r=0.138$, $P<0.05$. Two reviewers: $Y=5.0+0.05X$, $r=0.208$, $P<0.005$. Three reviewers: $Y=3.6+0.07X$, $r=0.249$, $P<0.0005$. Difference between the lines for one and three reviewers was significant (covariance analysis, $P<0.05$).

years for 183 original papers published between April 1998 and September 1999. All manuscripts had a reviewer's score based on three reports. For each paper the 3 years period (citation window) started in the month of publication and ended 36 months later. This procedure was applied because calendar years of citation (as used for calculation of the impact factor) have the disadvantage that January issues have a much longer period between their publication date and their citation window than December issues. The reviewer's priority scale indicates the number of high priority recommendations per paper divided by the number of reviewers ($\times 100$).

Fig. 2 shows that the priority recommendation of a single reviewer is positively correlated with future citation (see legend of Fig. 2 for numerical details). The combination of the scores of two reviewers predicted citation even better. The best result was obtained with the combined advice of three reviewers. Covariance analysis indicated that the differences between the lines for one and two reviewers and two or three reviewers were not significant. However, the difference between the lines for one and three reviewers was significant (covariance analysis, $P <$ 0.05). To our knowledge, this is the first quantitative evidence for the demonstration of the ability of reviewers to predict citations. A rigorous policy to refrain from publication of papers with a low priority indication may therefore potentially increase the impact factor. Note that the citation period was 36 months for all papers as opposed to the calculation of the impact factor with a citation window of only 12 months which, moreover, has a different time lag to the publication data of the issues which make up the impact factor (see Ref. [16]).

4. Can editors predict citation ?

Fig. 3 shows that the relationship between editor's rating (dotted line) and the reviewer's priority based on three reviewers (solid line) are equally potent in the prediction of citations (data on three reviewers from Fig. 2 depicted again in Fig. 3 for comparison). Interestingly, there was no longer a significant correlation between editor's rating and reviewer's priority in this subset of 183 published papers, which form a selected subgroup from the larger group of all submitted 3444 manuscripts in Fig. 1, where this relation was highly significant. Despite this, a combination of reviewer's priority and editor' rating had the strongest correlation with future citation (dashed line). We conclude from this that editors and reviewers both are capable in predicting citation, but probably they recognize different aspects of scientific quality. Also, these data show that the peer review process with external reviewers is of great benefit for the selection of frequently cited papers. With this knowledge we addressed whether and to what extent bias exists in the peer review process.
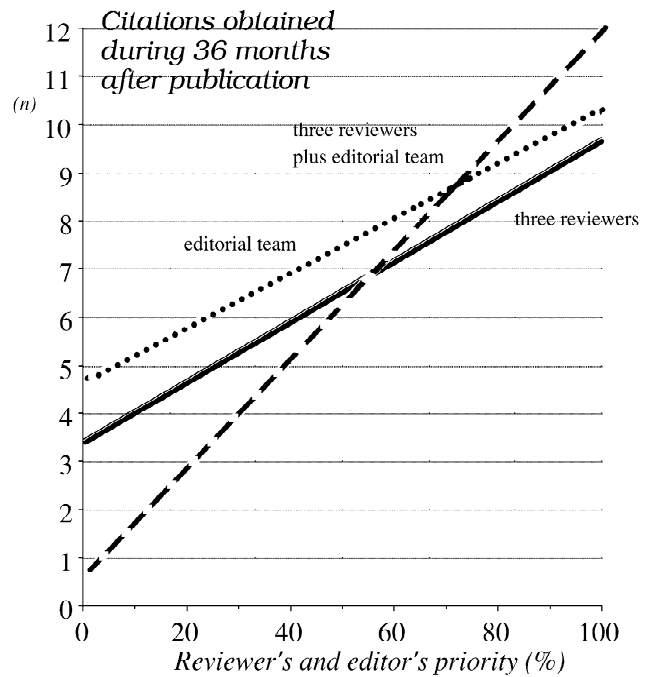

Fig. 3. Relationship between the priority recommendation of three reviewers and of the rating of the editors plus a combined rating of both on future citation over a period, which was exactly 3 years for each individual paper from a series of 183 papers published between April 1998 and September 1999 and cited between April 1998 and March 2001 (for the papers published in April 1998) and between September 1999 and August 2002 (for the papers published in September 1999). Three reviewers: $Y=3.6+0.07$ X, $r=0.249$, $P<0.0005$. Editorial team: $Y=4.7+0.06X$, $r=0.245$, $P<0.0005$. Combined reviewers+editor's score: $Y=0.7+0.12X$, $r=0.338$, $P<0.0005$.

5. Does bias exist for individual reviewers?

Siegelman [8] has analyzed the individual behaviour of reviewers. Only a very small minority of reviewers either assigns extremely high ratings to manuscripts ('zealots') or extremely low ratings ('assassins'). This was based on material from the journal *Radiology* (660 reviewers with at least 10 reports per person). These reviewers had been ranking manuscripts from 1 till 9. The percentage of reviewers with extreme ratings was about 1% at both sides of the spectrum.

In our analysis there were 334 reviewers with more than 10 reports. On average, they produced 14 reports. We scored their overall priority score and recalculated those as if each reviewer had produced 14 reports. Our reviewers were asked to choose between high (top 20%) priority and low priority. Fig. 4 shows that the observed distribution deviates substantially from the expected—calculated—distribution. Thus, based on an overall occurrence of 60% low priority and 30% high priority, there would be a chance of 0.08% (none expected in the group of 334 reviewers) to find 14 consecutive low priorities (bin 0–3%) based on binomial distribution calculations. However, we found
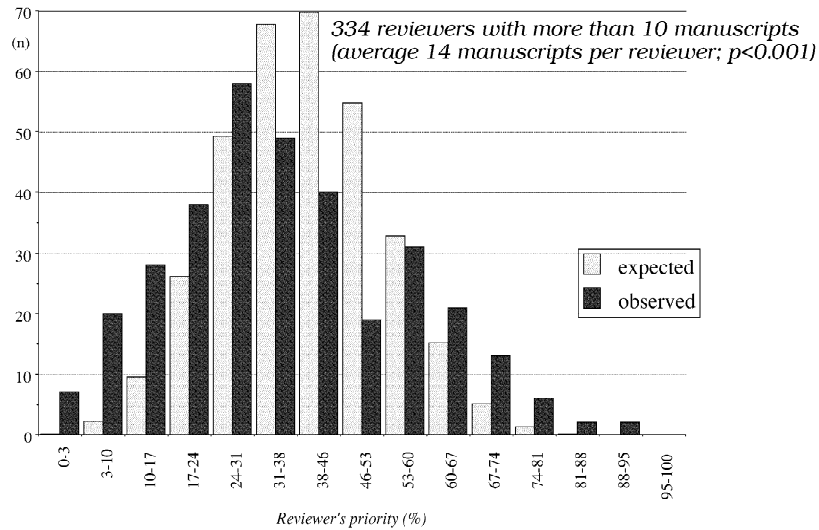
Fig. 4. A total of 334 reviewers delivered more than 10 reports. For each reviewer the individual priority score was calculated. A score of 0% indicated that the reviewer assigned a low priority to all manuscripts. A score of 100% meant that all manuscripts received a high priority. On the average, each of these reviewers produced 14 reports. A binomial calculation was performed on the data. The bin at the left (0–3%) indicates 14 low priorities. The next bin (3–10%) indicates 13 low priorities and one high priority and so on. The ordinate depicts the number of reviewers expected and observed in the respective bins. Within the first four left bins there were over 15% reviewers more than expected. See text for further explanation.

seven individuals within this bin. The chance to have 13 low priorities and one high priority (bin 3–10%) would amount to 0.73% or two individuals. Instead we observed 20 individuals. Within the first four bins (at the left) the expected cumulative percentage would be 12.43% with 41 individuals. Instead we found 93 individuals. Overall, 52 out of 334 reviewers (15.6%) tended to assign more low priorities to manuscripts than would be expected. This is substantially more than in the previous analysis of Siegelman [8].

6. Geographical aspects

We next hypothesized that if bias is present, it might relate to the country of origin of either authors or reviewers or both. We calculated the number of reviewer–manuscript interactions from the database with 3444 manuscripts. We omitted all manuscripts with authors with an affiliation from more than one country. This led to 8313 reviewer–manuscript interactions. Next, we selected the countries with more than 100 interactions both at the reviewer and the manuscript side. Some countries fulfilled one of the criteria, but not the other (reviewer: Belgium, Switzerland; manuscripts: Austria, Spain, Taiwan). Fig. 5 shows the countries grouped in the order of averaged reviewer's priority scores. The average priority score for all 8313 interactions was 35.07% and this value was set at 100% (dashed line). The reviewers histograms depict the priority score of the reviewers from a specific country on the manuscripts of all countries including the own country. The manuscript histograms depict the priority scores of the
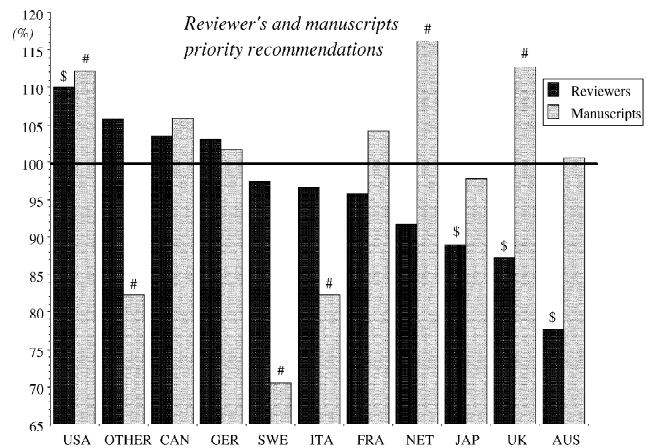

Fig. 5. The 3444 manuscripts received in general three reviewer's reports. The manuscripts with authors from more than one country were excluded. This led to 8313 reviewer–manuscript interactions. The overall priority score of those 8313 interactions was 35.07% and this value was set at 100% (dashed line). We selected those countries with more than 100 interactions both at the reviewer and manuscript side (see text for details). The other countries were grouped as 'other'. The countries are grouped along the abscissa according to their reviewer's priority scores. The reviewers histograms depict the priority score of the reviewers from a specific country on the manuscripts of all countries including the own country. The manuscript histograms depict the priority scores of the manuscripts from a specific country as assigned by the reviewers from all countries including the own country. A significantly higher reviewer score was observed for the USA and significantly lower scores were found for Japan, the United Kingdom (UK) and Australia (AUS). The negative score for The Netherlands (NET) was borderline significant. Manuscripts from The Netherlands, the United Kingdom and the USA scored significantly higher than the average, whereas the opposite was observed for manuscripts from Italy (ITA), 'other countries' (OTHER) and Sweden (SWE). See text for statistical details.

manuscripts from a specific country as assigned by the reviewers from all countries including the own country.

Reviewers from the USA assigned significantly higher priority to manuscripts (ANOVA, $P<0.0005$). The opposite was observed for reviewers from Japan ($P<0.025$), the United Kingdom ($P<0.001$) and Australia ($P<0.025$). Reviewers from The Netherlands also assigned relatively low priority scores (borderline significant).

Manuscripts from The Netherlands ($P<0.01$), the United Kingdom ($P<0.005$) and the USA ($P<0.0005$) received significantly higher priority ratings. Manuscripts from Italy ($P<0.01$), 'other countries' ($P<0.005$) and Sweden ($P<0.005$) scored significantly lower than the average.

These data do not necessarily demonstrate bias, because it cannot be excluded that manuscripts from one country are better than from other countries. Of course, this cannot easily explain the differences between reviewer's priority ratings (Fig. 5).

7. Geographical bias

We divided the 8313 reviewer–manuscript interactions described in Fig. 5 over 842 'matches' (reviewer and manuscript originated from the same country, 10%) and 7471 'non-matches' (90%). Fig. 6 shows that in case of a match there was a 23.6% higher priority score than when there was no-match (ANOVA, $P<0.0005$). After removal of all USA–USA interactions this 'over-rating' was still 25.3% (ANOVA, $P<0.0005$; data not shown). We next explored these relations for several countries, the choice of which was limited by the number of available interactions allowing statistical analysis.

7.1. Italy

Fig. 5 shows that Italian reviewers rate manuscripts not different from the average reviewer, but that Italian manuscripts receive a significantly lower score than average (82.3%, $P<0.01$). We therefore analyzed the evaluation of Italian manuscripts by Italian (Fig. 7. left bar) and non-Italian reviewers (second bar). Likewise, we analyzed non-Italian manuscripts (Fig. 7, right two bars). Fig. 7 shows that Italian reviewers tend to rate manuscripts from their own country even lower than reviewers from other countries (left two bars). These scores were 51.8 vs. 83.3% (ns). Dividing these numbers pointed to an 'Italian–Italian bias' of 0.62. In order to be able to judge whether this is the result of geographical bias or simply the result of a 'more demanding' Italian reviewer, we need to compare this ratio (0.62) with the ratio between the right two bars in Fig. 7 concerning non-Italian manuscripts (97.9% for Italian reviewers and 100.9% for non-Italian reviewers). This 'Italian–non-Italian bias was 0.97. If we correct the 'Italian–Italian bias' of 0.62 for the fact that Italian reviewers also slightly underrate non-Italian manuscripts (0.97), the overall 'nationality index' for Italian reviewers is 0.64, a measure of geographical bias (in this case negative geographical bias, no bias indicated by an index of 1.0).

No reviewers rated Italian manuscripts as low as Italian reviewers, although German reviewers came close. Highest ratings were received from French reviewers. The same calculations were performed for other countries.
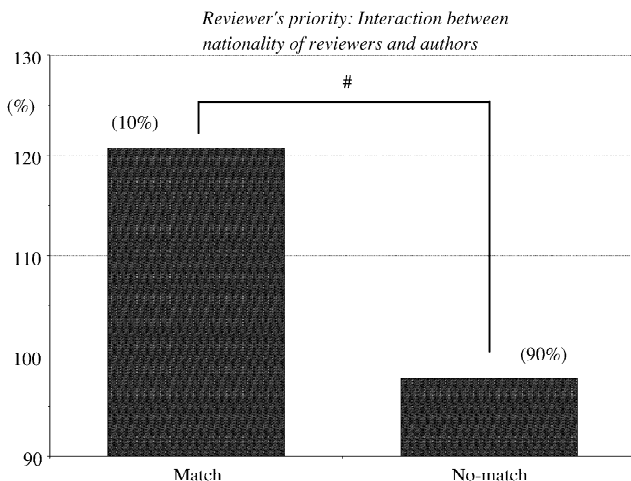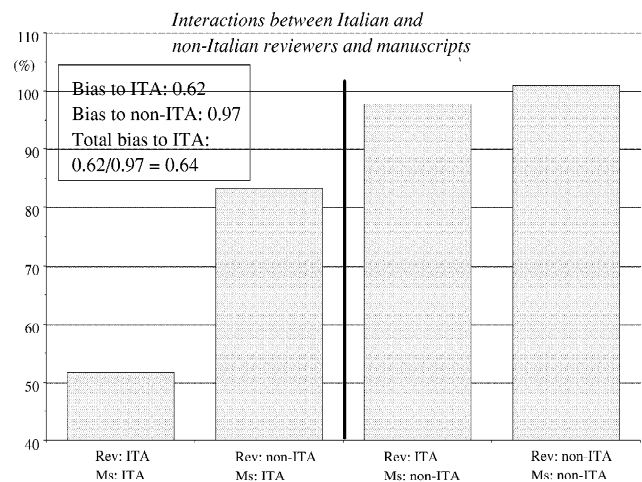

Fig. 6. There was a match between the country of the reviewer and of the authors of the manuscript in 842 cases out of 8,313 reviewer–manuscript interactions (10%). The 'over-rating' of these manuscripts was 23.6%, because the priority was 120.7% when there was a match and 97.7% when there was no match (ANOVA, $P<0.0005$, F value 23.96). When USA–USA interactions were excluded, the statistics were 25.3% 'over-rating' with 118.3% priority rating when there was a match and 94.4% when there was no match (ANOVA, $P<0.0005$, F value 481.58).

Fig. 7. Priority assignments of Italian and non-Italian reviewers to Italian and non-Italian manuscripts. Reviewer's bias was calculated for Italian manuscripts (0.62) and for non-Italian manuscripts (0.97). By dividing both ratios, the 'nationality index' was calculated. It was 0.64 for Italian reviewers.

7.2. USA

Fig. 5 indicates that American reviewers rate manuscripts significantly higher (110%, $P<0.0005$) than non-American reviewers. Also, American manuscripts received significantly higher priority ratings (112.2%, $P<0.0005$). In contrast to a previous study of Link [15], Fig. 8 shows that overrating of American reviewers compared to non-American reviewers applies as much to American manuscripts as it does to non-American manuscripts, giving rise to an overall 'nationality index' for American reviewers of 1.01.

No reviewers ranked American manuscripts as low as British reviewers, whereas the highest rankings came from Italian reviewers.

7.3. Japan

Fig. 5 shows that Japanese reviewers rate manuscripts significantly lower than non-Japanese reviewers (89.0%, $P<0.025$). Japanese manuscripts received average ratings (97.9%). Fig. 9 shows that the underrating of Japanese reviewers compared to non-Japanese reviewers was significant for non-Japanese manuscripts (right two bars), as it was for all manuscripts including the Japanese manuscripts (Fig. 5). The specific underrating of Japanese manuscripts by Japanese reviewers compared to non-Japanese reviewers (Fig. 9, left two bars) was borderline significant. Thus, the tendency for Japanese reviewers to underrate manuscripts was similar for Japanese and non-Japanese manuscripts, giving rise to an overall 'nationality-index' for Japanese reviewers of 1.04.

Lowest ratings were received from Australian and

Fig. 9. Priority assignments of Japanese and non-Japanese reviewers to Japanese and non-Japanese manuscripts. Reviewer's bias was calculated for Japanese manuscripts (0.91) and for non-Japanese manuscripts (0.88). By dividing both ratios, the 'nationality index' was calculated. It was 1.04 for Japanese reviewers.

Swedish reviewers, whereas the highest ratings came from Canadian and American reviewers.

7.4. United Kingdom

Fig. 5 indicates that British reviewers rate manuscripts significantly lower than non-British reviewers (87.1%, $P<0.001$). British manuscripts received significantly higher ratings than average (112.8%, $P<0.005$). Fig. 10 shows that British reviewers assign a 143.9% rating to British manuscripts and 80.4% to non-British manuscripts. Non-
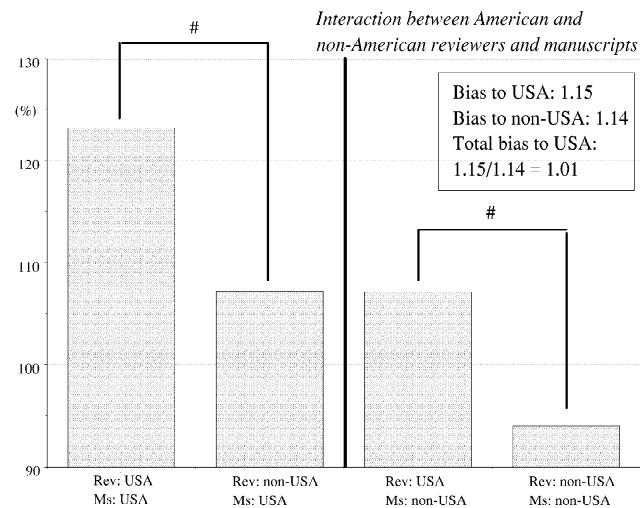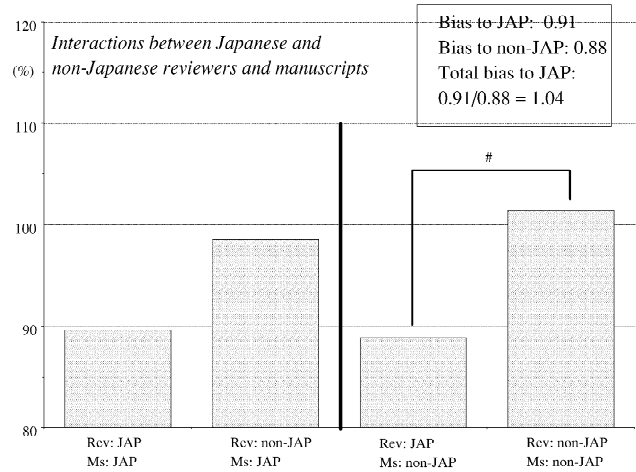
Fig. 8. Priority assignments of American and non-American reviewers to American and non-American manuscripts. Reviewer's bias was calculated for American manuscripts (1.15) and for non-American manuscripts (1.14). By dividing both ratios, the 'nationality index' was calculated. It was 1.01 for American reviewers.
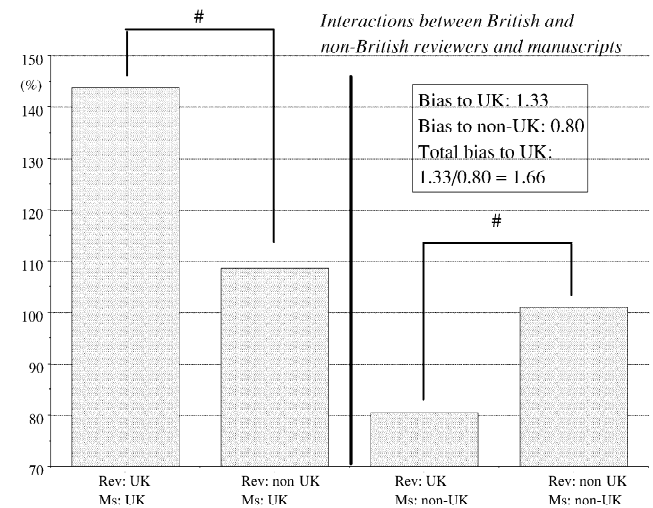
Fig. 10. Priority assignments of British and non-British reviewers to British and non-British manuscripts. Reviewer's bias was calculated for British manuscripts (1.33) and for non-British manuscripts (0.80). By dividing both ratios, the 'nationality index' was calculated. It was 1.67 for British reviewers.

British reviewers rate British manuscripts by 108.6% and non-British manuscripts by 100.9%. The ratio of the priority of British reviewers to non-British reviewers on British manuscripts was 1.33 ($P<0.01$). That same ratio on non-British manuscripts was 0.80 ($P<0.001$), leading to a 'nationality index' of 1.67 (positive geographical bias).

British manuscripts received the lowest ratings from Australian and Canadian reviewers and the highest from British reviewers.

7.5. France

Fig. 5 shows that French reviewers (95.8%) as well as French manuscripts (104.2%) scored priority ratings that were not different from the average. Fig. 11 shows that French reviewers assigned a 165.1% rating to French manuscripts, although the rating from non-French reviewers was 101.7%. The difference was significant ($P<0.05$). Fig. 11 also shows that French reviewers rated non-French manuscripts by 92.7% compared to 100.1% by non-French reviewers. These data gave rise to a French 'nationality index' of 1.75.

French manuscripts received the lowest ratings from Australian reviewers and the highest ratings from French reviewers.

7.6. Other countries

The 'nationality index' for Sweden was 1.03, for Germany 1.04 (see also Ref. [17]), for Canada 1.20 and for The Netherlands 1.21.
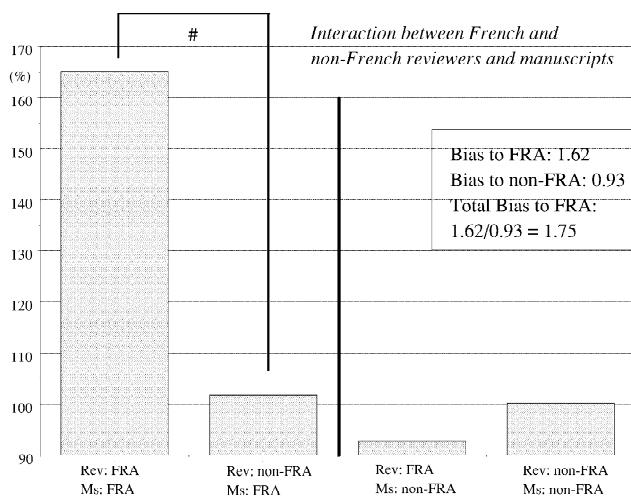

Fig. 11. Priority assignments of French and non-French reviewers to French and non-French manuscripts. Reviewer's bias was calculated for French manuscripts (1.62) and for non-French manuscripts (0.93). By dividing both ratios, the 'nationality index' was calculated. It was 1.75 for French reviewers.

8. Conclusions

We conclude that the peer review process is valuable in selecting highly cited papers and therefore cannot be dismissed. In spite of this, we unequivocally demonstrate the presence of (positive as well as negative) geographical bias in the interaction between reviewers and authors.

We wish to underscore that the peer review process is not merely employed by editors as a selection method, but serves many other purposes as well. It primarily assists in improving the quality of the submitted manuscripts, whether the manuscripts are accepted or rejected. The above demonstration of geographical bias should be seen in the light that peer review is more than ticking a box with 'high' or 'low' priority. The unrestricted exchange of thought and criticism is at the root of the scientific process and peer review therefore is part of this process. However, on the basis of our data it may be argued that anonymity of peer review should be lifted because this may decrease both personal as well as geographical bias, but probably other types of bias (gender-, career-, competitive-) as well.

9. Summary of results

- The concordance between editor's and reviewer's ratings is significant for submitted papers, but not sufficient to allow in-house rejection.
- Both editor's ratings and reviewer's ratings predict future citation.
- Ratings of editors and reviewers are not significantly correlated for accepted manuscripts. Despite this the highest ratings from editors and reviewers have the strongest predicitive power for future citation.
- Manuscripts receive significantly higher priority ratings when reviewers and authors originate from the same country.
- American reviewers rate manuscripts higher than non-American reviewers regardless which manuscripts are involved.
- British and French reviewers assign significantly higher priority ratings to manuscripts from their own country than from other countries.

References

[1] Coronel R, Opthof T. The role of the reviewer in editorial decision-making. Cardiovasc Res 1999;43:261–264.

[2] Kronick DA. Peer review in the 18th-century scientific journalism. J Am Med Assoc 1990;263:1321–1322.

[3] Burnham JC. The evolution of editorial peer review. J Am Med Assoc 1990;263:1323–1329.

[4] Guarding the guardians: research on editorial peer review: selected proceedings from the first international congress on peer review in biomedical publication. J. Am. Med. Assoc. 1990; 263: 1317-1441 (theme issue).

[5] The Second International Congress on Peer Review in Biomedical Publication. J. Am. Med. Assoc. 1994; 272: 91-173 (theme issue).

[6] The Third International Congress on Peer Review in Biomedical Publication. J. Am. Med. Assoc. 1998; 280: 213-302 (theme issue).

[7] The Fourth International Congress on Peer Review in Biomedical Publication. J. Am. Med. Assoc. 2002; 287: 2759-2871 (theme issue).

[8] Siegelman SS. Assassins and zealots: variations in peer review. Radiology 1991;178:637–642.

[9] Scharschmidt BF, DeAmicis A, Bacchetti P, Held MJ. Chance, concurrence, and clustering. Analysis of reviewer's recommendations on 1,000 submissions to The Journal of Clinical Investigation. J Clin Invest 1994;93:1877–1880.

[10] Peters DP, Ceci SJ. Peer-review practices of psychological journals: the fate of submitted articles, submitted again. Behav Brain Sci 1982;5:187–195.

[11] Wilson JD. Peer review and publication. J Clin Invest 1978;61:1697–1701.

[12] Opthof T, Furstner F, Van Geer M, Coronel R. Regrets or no regrets? No regrets! The fate of rejected manuscripts. Cardiovasc Res 2000;45:255–258.

[13] Gilbert JR, Williams ES, Lundberg GD. Is there gender bias in JAMA's peer review process? J Am Med Assoc 1994;272:139–142.

[14] Wenneras C, Wold A. Nepotism and sexism in peer-review. Nature 1997;387:341–343.

[15] Link AM. US and non-US submissions. An analysis of reviewer bias. J Am Med Assoc 1998;280:246–247.

[16] Opthof T. Sense and nonsense about the impact factor. Cardiovasc Res 1997;33:1–7.

[17] Opthof T, Coronel R, Janse MJ. Submissions, impact factor, reviewer's recommendations and geographical bias within the peer review system (1997–2002). Focus on Germany. Cardiovasc Res 2002;55:215–219.