

EXPERT COMMENTARY

Does the Journal Peer Review Select the “Best” from the Work Submitted? The State of Empirical Research

Lutz Bornmann

Professorship for Social Psychology and Research on Higher Education, Department of Humanities, Social and Political Sciences, ETH Zurich, Zähringerstr. 24, CH-8092 Zurich, Switzerland

Abstract

The goal for the peer review of manuscripts is usually to select the “best” from the work submitted. In the investigation of the predictive validity of the peer review process, the question of whether this goal will actually be achieved arises, that is, whether indeed the “best” manuscripts are published. This review of the studies published on the predictive validity of peer review describes the state of empirical research on this topic. All studies confirmed that the editorial decisions (acceptance or rejection) for the various journals appear to reflect a rather high degree of predictive validity, if citation counts are employed as validity criteria. The studies confirm that peer review represents a quality filter and works as an instrument for the self-regulation of science.

Keywords

Journal peer review, Predictive validity, Editorial decision

1. Introduction

The goal for peer review of manuscripts is usually to select the “best” from the work submitted [1]. In the investigation of the predictive validity of the peer review process, the question of whether this goal will actually be achieved arises, that is, whether indeed the “best” manuscripts are published. The validity of judgments in peer review is often questioned. For example, the former editor of the journal *Lancet*, Sir Theodore Fox [2], writes on the validity of editorial decisions: “When I divide the week’s contributions into two piles—one that we are going to publish and the other that we are going to return—I wonder whether it would make any real difference to the journal or its readers if I exchanged one pile for another” (p. 8). The selection function is considered to be a difficult research topic to investigate. According to Jayasinghe, Marsh, and Bond [3] and Figueredo [4], there exists no mathematical formula or uniform definition of what makes a manuscript “worthy of publication.”

2. Research on the Predictive Validity of Journal Peer Review

One first step in testing the predictive validity of a peer review process comes from investigating the fate of rejected manuscripts: “A rejection usually does not kill a paper...a rejected paper usually finds life at another journal” ([5], p. 177). Already in the 1980s, Abelson [6] reported that almost all of the manuscripts rejected by

the journal *Science* were published later in other journals. For manuscripts rejected by the journal *Angewandte Chemie International Edition* (AC-IE) in the year 1984, Daniel [7] determined a percentage of 71%; for manuscripts rejected by this journal at the beginning of the year 2000, Bornmann and Daniel [8,9] found that 95% were published later elsewhere. (Interestingly, the follow-up study could determine that no alterations or only minor alterations were made to approximately three-quarters of the rejected manuscripts for publication elsewhere.) Other studies on the fate of manuscripts rejected by a journal report percentages ranging from 28% to 85% [10]. Taken altogether, these studies demonstrate that in the peer review process of one and the same manuscript submitted various times, reviewers and editors arrive at different judgments: Manuscripts that are rejected by a journal (through a peer review process) are then accepted by another journal (through a peer review process). This finding indicates that manuscript review is not only based on generally valid quality criteria that a scientific work can fulfill (acceptance) or not fulfill (rejection); the review (or the outcome of the review) also seems to be dependent upon the local conditions under which the peer review process at the individual journals takes place.

Following recommendations, such as those of Harnad [11], that “peer review [has] to be evaluated objectively (i.e. via metrics)” (p. 103), the second step in the research on the predictive validity of journal peer review consists of gauging the quality of journals that accepted previously

rejected manuscripts. According to Jennings [12], “there is a hierarchy of journals. At the apex of the (power law-shaped?) pyramid stand the most prestigious multidisciplinary journals; below them is a middle tier of good discipline-specific journals with varying degrees of selectivity and specialization; and propping up the base lies a large and heterogeneous collection of journals whose purviews are narrow, regional or merely unselective.” In her literature review covering research on journal peer review, Weller [10] cites five studies [13-17] which have ranked the quality of rejecting and later accepting journals mostly by means of the Journal Impact Factors (JIF, provided by Thomson Reuters, Philadelphia, PA, in the Journal Citation Reports, JCR). The JIF is the average number of times papers from the journal published in the past 2 years (e.g. 2005 and 2006) have been cited in the JCR year (e.g., 2007) [18,19].

Six further studies, which are not included in the literature review by Weller [10], have been published by Bornmann and Daniel [8,9], Daniel [7], Lock [20], McDonald, Cloft, and Kallmes [21], Opthof, Furstner, van Geer, and Coronel [22], and Ray, Berkwits, and Davidoff [23]. In the total of 11 studies, between 0% [7] and 70% [17] of the rejected manuscripts in a higher quality journal could be researched. The results of these studies show accordingly “that authors do not necessarily move from ‘leading’ journals to less prestigious journals after a rejection” ([10], p. 68). Authors seem to select a journal for a rejected manuscript based on the quality of the rejecting journal and the availability of additional high(er)-impact journals: The higher the quality ranking of the rejecting journal, the lower the chance that a rejected manuscript will appear in another journal ranked as higher quality.

A third and most important step for the investigation of the predictive validity of peer review consists of comparing the impact of papers accepted or rejected (but published elsewhere) in journal peer review. As the number of citations of a publication reflects its international impact [24] and because of lack of other operationalizable indicators, it is a common approach in peer review research to evaluate the success of the process on the basis of citation counts. Citation counts are attractive raw data for the evaluation of research output: They are “unobtrusive measures that do not require the cooperation of a respondent and do not themselves contaminate the response (i.e., they are non-reactive)” ([25], p. 84). Although citations have been a controversial measure of both quality and scientific progress [26], they are still accepted as a measure of scientific impact, and thus as a partial aspect of scientific quality [27].

Scientific judgments on manuscripts are said to show predictive validity in peer review research, if the citation counts of manuscripts accepted for publication and manu-

scripts rejected by a journal but then published elsewhere differ statistically significantly. Up until now only a few studies have conducted analyses which examine citation counts from individual papers as the basis for assessing predictive validity in peer review. Research in this area is extremely labor intensive, since a validity test requires information and citation counts regarding the fate of rejected manuscripts [28]. The editor of the *Journal of Clinical Investigation* [29] has undertaken his own investigation into the question of predictive validity. Daniel [7] and Bornmann and Daniel [8,9] investigated the peer review process of AC-IE, and Opthof *et al.* [22] did the same for *Cardiovascular Research*. McDonald, Cloft, and Kallmes [30] examined the predictive validity of the editorial decisions for the *American Journal of Neuroradiology*. All five studies confirmed that the editorial decisions (acceptance or rejection) for the various journals appear to reflect a rather high degree of predictive validity, if citation counts are employed as validity criteria. The studies confirm that peer review represents a quality filter and works as an instrument for the self-regulation of science.

3. Conclusions

Against the studies, which have investigated the predictive validity of selection decisions in peer review on the basis of bibliometric data, a few critical points have been articulated. As shown, five studies on journal peer review have found a high degree of predictive validity through the comparison of mean citation rates for accepted manuscripts and rejected manuscripts published elsewhere. Cicchetti [31] has raised an argument against this form of validity test, pointing out that papers accepted by journals (that studies have investigated, e.g., AC-IE) may have been cited on average more frequently than those published elsewhere simply because they appeared in journals with a high JIF. Higher citation rates are not necessarily the result of a paper’s superior scientific quality (which is reflected in its acceptance for publication); instead, they may just show the higher impact or higher visibility of a journal. According to the results of Seglen [32], however, the citation counts of articles do not seem to be detectably influenced by the status of the journals in which they are published. Against the criticism of Cicchetti [31], this form of validity test for journal peer review decisions should thus enable valid results.

The results of most studies on the predictive validity of journal peer review base on statistical methods, which strictly speaking should not be applied to bibliometric data. For example, citation impact differences between accepted manuscripts and manuscripts that were rejected but published elsewhere were determined on the basis of arithmetic means. As a rule, the distribution of citation counts for a larger number of publications is skewed to the right according to a power law [33]. In the face

of non-normal distributed citation data, the arithmetic mean value is not appropriate as it can give a distorted picture of the kind of distribution and "it is a rather crude statistic" ([33], p. 2).

Comparisons drawn between groups of papers as to research performance are according to Bornmann, Mutz, Neuhaus, *et al.* [34] valid only if (1) the scientific impact of the groups are looked at by using box plots, Lorenz curves, and Gini coefficients to represent distribution characteristics of data (in other words, going beyond the usual arithmetic mean value), (2) different reference standards are used to assess the impact of the groups and the appropriateness of the reference standards undergoes critical examination, and (3) the comparative analysis of the citation counts for publications takes into consideration that in statistical analysis, citations are a function of many influential factors besides scientific quality. Among others, the influential factors include number of co-authors, location of the authors, the prestige, language, and availability of the publishing journal, and the size of the citation window [26]. By including these factors in the statistical analysis, it becomes possible to establish the adjusted covariation between selection decisions and citation counts.

A general weakness of the research on the predictive validity of peer review is above all the lack of studies [10]. Further comprehensive research is still lacking.

References

1. R. Smith. "Peer review: A flawed process at the heart of science and journals," *Journal of the Royal Society Med*, vol. 99, pp. 178-82, Apr. 2006.
2. T. Fox. *Crisis in communication: The functions and future of medical publication*. London, UK: Athlone Press; 1965.
3. U.W. Jayasinghe, *et al.* "Peer review in the funding of research in higher education: The Australian experience," *Edu Eval Policy Anal*, vol. 23, pp. 343-6, 2001.
4. E. Figueredo. "The numerical equivalence between the impact factor of journals and the quality of the articles," *J Am Soc Inform Sci Tech*, vol. 57, p. 1561, 2006.
5. J.S. Gans, and G.B. Shepherd. "How are the mighty fallen - rejected classic articles by leading economists," *J Eco Perspect*, vol. 8, pp. 165-79, WIN 1994.
6. P.H. Abelson. "Scientific communication," *Science*, vol. 209, pp. 60-2, 1980.
7. H.D. Daniel. *Guardians of science. Fairness and reliability of peer review*. Weinheim, Germany: Wiley-VCH. Wiley Interscience, 1993.
8. L. Bornmann, and H.D. Daniel. "The effectiveness of the peer review process: Inter-referee agreement and predictive validity of manuscript refereeing at *Angewandte Chemie*," *Angewandte Chemie Inter ed*, vol. 47, pp. 7173-8, 2008.
9. L. Bornmann, and H.D. Daniel. "Selecting manuscripts for a high impact journal through peer review: A citation analysis of Communications that were accepted by *Angewandte Chemie International Edition*, or rejected but published elsewhere," *J Am Soc Inform Sci Tech*, vol. 59, pp. 1841-52, 2008.
10. A.C. Weller. *Editorial peer review: Its strengths and weaknesses*. Medford, NJ, USA: Information Today, Inc., 2002.
11. S. Hamad. "Validating research performance metrics against peer rankings," *Ethics Sci Environ Pol*, vol. 8, pp. 103-107, 2008.
12. C.G. Jennings. *Quality and value: The true purpose of peer review. What you can't measure, you can't manage: The need for quantitative indicators in peer review*. Available from: <http://www.nature.com/nature/peerreview/debate/nature05032.html> [last cited on 2006 Jul 6].
13. F.S. Chew. "Fate of manuscripts rejected for publication in the *AJR*," *AJR Am J Roentgenol*, vol. 156, pp. 627-32, Mar 1991.
14. B. Cronin, and G. McKenzie. "The trajectory of rejection," *J Document*, vol. 48, pp. 310-7, Sep. 1992.
15. N. Whitman, and S. Eyre. "The pattern of publishing previously rejected articles in selected journals," *Family Med*, vol. 17, pp. 26-8, 1985.
16. A.C. Weller. "Editorial peer review: A comparison of authors publishing in two groups of US medical journals," *Bull Medical Library Assoc*, vol. 84, pp. 359-66, Jul 1996.
17. M.D. Gordon. "How authors select journals - a test of the reward maximization model of submission behavior," *Social Studies Sci*, vol. 14, pp. 27-43, 1984.
18. L. Bornmann, *et al.* "Citation environment of *Angewandte Chemie*," *CHIMIA*, vol. 61, pp. 104-9, 2007.
19. L. Bornmann, and H.D. Daniel. "The state of h index research. Is the h index the ideal way to measure research performance?," *EMBO Reports*, vol. 10, pp. 2-6, 2009.
20. S. Lock. *A difficult balance: Editorial peer review in medicine*. Philadelphia, PA, USA: ISI Press, 1985.
21. R.J. McDonald, *et al.* "Fate of submitted manuscripts rejected from the *American Journal of Neuroradiology*: Outcomes and commentary," *Am J Neuroradio*, vol. 28, pp. 1430-4, Sep. 2007.
22. T. Opthof, *et al.* "Regrets or no regrets? No regrets! The fate of rejected manuscripts," *Cardiovas Res*, vol. 45, pp. 255-8, Jan. 1, 2000.
23. J. Ray, *et al.* "The fate of manuscripts rejected by a general medical journal," *Am J Med*, vol. 109, pp. 131-5, Aug. 1, 2000.
24. M.J. Kumar. "Evaluating scientists: Citations, impact factor, h-Index, online page hits and what else," *IETE Technical Review*, vol. 26, pp. 165-8, May-Jun. 2009.
25. L.C. Smith. "Citation analysis," *Library Trends*, vol. 30, pp. 83-106, 1981.
26. L. Bornmann, and H.D. Daniel. "What do citation counts measure? A review of studies on citing behavior," *J Document*, vol. 64, pp. 45-80, 2008.
27. B.R. Martin, and J. Irvine. "Assessing basic research - some partial indicators of scientific progress in radio astronomy," *Res Policy*, vol. 12, pp. 61-90, 1983.
28. R.F. Bornstein. "The predictive validity of peer-review: A neglected issue," *Behavioral Brain Sci*, vol. 14, pp. 138-9, Mar. 1991.
29. J.D. Wilson. "Peer review and publication," *J Clinical Invest*, vol. 61, pp. 1697-701, 1978.
30. R.J. McDonald, *et al.* "Fate of manuscripts previously rejected by the *American Journal of Neuroradiology*: A follow-up analysis," *Am J Neuroradio*, vol. 30, pp. 253-6, Feb 2009.
31. D.V. Cicchetti. "Guardians of science: Fairness and reliability of peer review," *J Clinical Experiment Neuropsych*, vol. 21, pp. 412-21, Jun. 1999.
32. P.O. Seglen. "Causal relationship between article citedness and journal impact," *J Am Society Inform Sci*, vol. 45, pp. 1-11, Jan. 1994.
33. Joint Committee on Quantitative Assessment of Research, "Citation statistics. A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)," International Mathematical Union (IMU), Berlin, Germany: 2008.
34. L. Bornmann, *et al.* "Use of citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results," *Ethics Sci Environ Pol*, vol. 8, pp. 93-102, 2008.

AUTHOR



Lutz Bornmann is a researcher at the Professorship for Social Psychology and Research on Higher Education of the ETH Zurich. Since the late 1990s, he has been working on issues in the promotion of young academics and scientists in the sciences and on quality assurance in higher education. He is a member of the editorial board of the *Journal of Informetrics* and

of the advisory editorial board of *EMBO Reports* (Nature Publishing group). Since 2004, he has published more than 40 papers in journals covered by Thomson Reuters with a total of nearly 400 citations. His *h* index amounts to 10 (the publication list and citation metrics are available on <http://www.researcherid.com/rid/A-3926-2008>). Thomson Reuters lists five of his papers in the Essential Science Indicators as highly cited papers (belonging to the top 1% of papers in the social sciences).

E-mail: bornmann@gess.ethz.ch

DOI: 10.4103/0256-4602.60162; Paper No TR 283_09; Copyright © 2009 by the IETE