Plasticity and Bias in Peer Assessment: Critical Perspectives from Bibliometrics

Ulf Sandström[1]

[1] *ulf.sandstrom@indek.kth.se*
Royal Institute of Technology, Dept Indek, SE-10044 Stockholm, Sweden

Abstract

This research paper in progress discusses some of the common criticisms of peer review: Costs and Robustness, Nepotism (conflict of interest), Sexism and Cognitive Bias. Attention is given to the fact that much of the research reported fails on a crucial point: The use of bibliometrics as a correlate for the grading and ranking done by granting or evaluation committees (ad hoc or standing committees).
The full paper will extend the analysis using data from a selection of finished projects and assessments. Results indicate that there are systemic problems regarding peer review: Firstly, the positive bias in university assessments based on ad hoc committees. Problems circulate around the absence of robust benchmarks and the ad hoc selection of experts. Secondly, the role of cognitive distance points to the power mechanisms in selection processes for finding relevant reviewers. Thirdly, the low levels of peer's performance (in bibliometric respect) indicate that selection of peers is no longer to search for the best possible peer, but instead, the pragmatic peer.

Introduction

For different reasons bibliometric experts are often defensive in relation to peer review. Typically, it is pointed out that bibliometrics cannot stand alone but should be used alongside with peer assessment. Despite the inconsistencies and shortcomings of occasional peer review, the majority of senior researchers trust collegial processes and, therefore, insists that peer assessment is the best alternative for identifying "quality" in science. Anecdotes and snapshots from the review process are often based on circular definitions like "you know quality when you see it" or "excellence smash you in the face". But, there is very little interest to explain why the grades given to female researchers who applied for grants from the Swedish MFR could change overnight to such an extent that success rates in 1996 increased with almost 30 per cent (from about 35 % to 48 %). Could it, maybe, be the effect of the first Wennerås and Wold report published early in 1995?
This paper concerns committee grant peer review and committee research assessments. Consequently, journal peer review is not discussed.

Peer Review Criticisms: notes to the literature

There are, of course, limitations to the bibliometric indicators as well, but nowadays bibliometricians have a better case as they can rely on normalized indicators and advanced statistical measures. In the following we will discuss some of most common criticisms of grant peer review.[18]
Reliability, Cost and Robustness: The main critical remark towards peer assessment focuses on low reliability (Cicchetti, 1991), although peers might have higher inter-reviewer agreement on the "best" and the "worst". In England another criticism has spurred a debate; the Research Assessments are creating a heavy economic burden due the huge costs and the low robustness, i.e. the inability to use peer review data to study shifts in quality over time (HEFCE, 1997). In order to have an international benchmark, there is always need for bibliometrics.

[18] Aksnes and Taxt (2004) report and discuss some of the "mistakes" done by peers in assessments of Norwegian research groups.

Nepotism: Due to lack of data there are seldom studies on conflicts of interest, but a few studies have shown the influence of having a "friend" in the committee (Wenneras & Wold, 1997; Sandström & Hällsten, 2008; see also Juznic et al. 2010). The former two studies used the same standard definition of "conflict of interest" based on self-reporting by the members of the committee and noted in the protocol from the review committee of the research council. With statistical methods it was possible to measure how many publication points or grading points that was credited to those with friends in the committee in relation to their bibliometric performance relative to other applicants. More studies are wanted on this aspect of peer review.
Sexism: Wenneras and Wold [W&W] (1997), in their classical study, showed that sexism could be a strong component that explains the result of a peer review process. The research council board had to resign soon after this was revealed. A new board was installed and with the assignment to report each year to the government on improvements in gender equality. Ten years later Sandström & Hällsten [S&H] (2008) did a follow-up study using the same methodology. They could show that the gender issue more or less was turned up-side-down. Male applicants had a disadvantage. This indicates plasticity in the peer review mechanism. Furthermore, another interesting result from the S&H study was the persistence of nepotism (conflict of interest). The friendship factor gave a bonus of about 15 per cent. The awareness of gender equality during the late 1990's in Sweden had obviously paid off and resulted in a general advantage for female applicants.
From one perspective this type of unfairness could be legitimized as female researchers in general seem to have fewer publications; that is, they publish fewer papers but their papers are more cited. As long as the review procedure appears to be based on the length of the publication list and not on the quality of papers it seems fair to have some sort of adjustment for female researchers.
Cognitive bias: Travis & Collins (1991) pioneered in putting this aspect along with other criticisms of peer review. Their study was based on direct observation of expert committee meetings. Although many others have attempted to replicate their research e.g. the psychologist Sven Hemlin, it must be concluded that it is very hard to get access for direct observation. Lamont (2009) evokes and opening up of the black box of peer review through interviews, and thereby she illustrates that emotions and informal criteria seem to play a significant role in the process. According to several investigations many applicants complain that reviewers are remote to their own field and do not understand the actual research question (Sandström et a. 1997; Wessely, 1998). At the same time we have to acknowledge Bourdieu's statement: People continuously manipulate the social reality and this reality exists largely in the discourse.
Whether inappropriate selection of reviewers is a common phenomenon is, of course, hard to tell, but from that question we can arrive at a similar, almost akin, question: May the reviewer selections make it harder for some applicants while others gain from that selection? In this paper we will show that this might be the case using data on cognitive distance among reviewers and applicants (c.f. Sandström, 2009).

The missing bibliometric component in studies on peer review

The interpretation of W&W differs a lot between the many scientific communities that are engaged in discussions over gender and grant-giving organizations. The importance of the study is should be underlined as it is one of few studies that use data CV together with all relevant bibliometric data. This is seldom the case as it is burdensome to put together publication data for a large selection of researchers. W&W (1997) used data from 112 individual cases and S&H (2008) had 280 cases. Obviously, the design of the W&W study has not been fully understood by several of the subsequent investigations.

As there were several investigations published on gender effects so called meta-analysis have been performed. One study (Bornmann et al., 2007) used 21 publications and 66 reviews and concluded that female researchers are at a disadvantage in peer review of grant applications. One year later, an analysis of a large dataset from the Australian Research Council found no evidence of gender bias in peer-reviewed grant funding. One conclusion of the latter study was that the result found by W&W study was contradicted.
Given the conflicting results between the two meta-studies, the two research teams joined forces to extend the original meta-analysis. The extended analysis (Marsh et al. 2009) indicated that the applicant's gender had no effect on funding and persisted in their criticism towards the W&W study although they had no data on the performance of researchers.
Why then is the bibliometric component of such an importance? The answer is basically that we have no information on whether the applicants are statistically representative of their respective groups (male-female) or not is unclear, and in all cases we have no information on whether there are any self-selection processes in action. It might be that the selection of female researchers is very exclusive and it could be the other way around for male researchers. If that is the case we would expect better bibliometric results from the former group and, in general, less good results from the male group. Therefore, there is need for a correlate to the committee decisions. Here, we assume that bibliometric data are more or less unbiased for each applicant and that full bibliometric data with relative citations scores will produce relevant correlates to the grading and ranking procedures of standing committees or ad hoc committees.
The Marsh et al. study (2009) drew conclusions on peer review without having a correlate and we find it very disturbing that they criticize the W&W study without having any possibility of presenting a comparable and relevant investigation as they did not have any bibliometric data at hand.
Another question, that has been raised from time to time, is whether the regression models used in the W&W study are relevant on not (Ceci & Williams, 2011). Using the original W&W data in a replication we find that this is not a problem worth mentioning.

Data and Method

As stated above, this paper aims at contributing to the ongoing debate on peer review from a bibliometrics. Five sets of data used in recent projects will be exploited (of which only the first is reported here):

- Peer judgement and bibliometric performance from a number of research assessment exercises at Swedish and Finnish Universities, (from Sweden: Uppsala, Lund, KTH, SLU, MiUN, ORU, JH and from Finland: Aalto University in Helsinki).
- Evaluation data from five area evaluations organized by the Swedish Research Council in 2001-2003 (chemical engineering, biotechnology, meteorology, plant science and theoretical chemistry).
- Data on cognitive distance among reviewers and applicants to the Strategic Foundations excellence centres in 2002 (c.f. Sandström et al. 2010)
- Data on the bibliometric performance by panel members of Swedish Council of Medicine (SCM).

Research Assessment Exercises

The Swedish RAE:s (and the Finnish) have produced evaluation data for almost 335 research units (≈ research groups). Peer judgements have been transformed to a unified grading score in five categories from Outstanding (5) to insufficient (1). In parallel all units have been scrutinized by bibliometric performance measures. Uppsala and Lund bibliometrics was done by the Leiden group. All the others were done with slightly different methods (Leydesdorff &

Opthof, 2010, Leydesdorff & Opthof, 2011; Waltman et al. 2011) by the author and his team. Figure 1 show the bibliometric performance (NCSf=field normalized citation score) as a distribution over citation classes. As expected we find that it approximates the normal distribution (c.f. van Raan, 2006) and we can put in grades according to bibliometric performance using a standard deviation of 1. This gives the thresholds for a five graded system just as the one used for peer assessments.

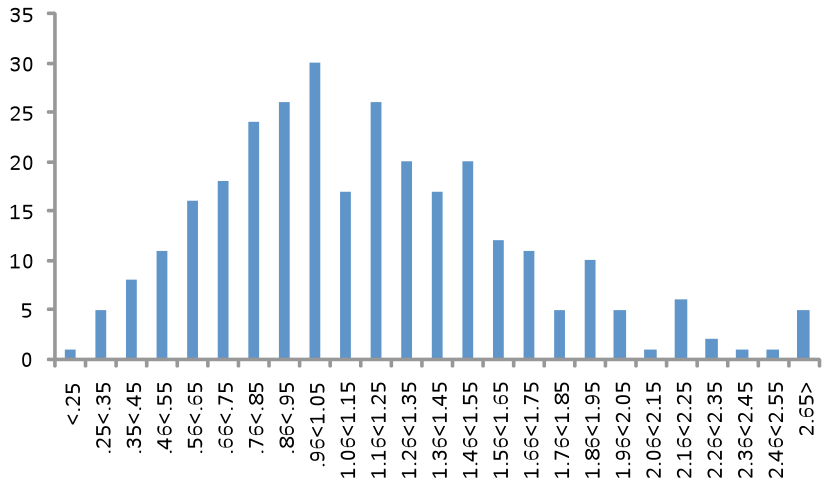

Figure 1. Distribution over citation classes (NCSf) for 298 units of assessment

When comparing the two different assessment methods – peers and metrics – we find that there is a considerable mismatch, se Table 1. Not more than 31 per cent of cases have an identical evaluation, and if we accept a peer assessment of +-1 we receive a figure of 73 per cent. Still, accepting quite a large variation we find that only three out of four cases show similarity between metrics and peer assessment. Obviously, one explanation to this result is the positive bias in peer assessments. The typical grade given by peers is "Excellent" (i.e. grade 4).

Table1. Comparison between peer assessment and metrics

	Field Normalized Bibliometric Performance					
Peer Assessment	1	2	3	4	5	Total
1	**4**	6	1			11
2	4	**11**	9	3	1	28
3	4	17	**36**	16	3	76
4	6	26	43	**22**	6	103
5	1	5	24	13	**10**	53
Total	**19**	**65**	**113**	**54**	**20**	**271**

Note: Comparison based on 271 units visible in ISI.

Conclusion: Three points for further discussion

The full paper will extend the analysis to the projects mentioned above (bullet points). Results indicate that there are systemic problems regarding peer review:
Firstly, the positive bias in university assessments based on ad hoc committees. Problems circulate around the absence of robust benchmarks and the ad hoc selection of experts.
Secondly, the role of cognitive distance points at the power mechanisms in the selection process for finding relevant reviewers.
Thirdly, the low levels of peer's performance (in bibliometric respect) indicate that selection of peers is no longer to search for the best possible peer, but instead, the pragmatic peer.

References

Aksnes DW & Taxt RE (2004) Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. Research Evaluation 13: 33-41.
Bornmann L (2007) Bias cut: women, it seems, often get a raw deal in science… *Nature* 445 (7127): 566.
Cicchetti DV (1991). The reliability of peer-review of manuscript and grant submissions: a cross-disciplinary investigation. Behavioral and Brain Science 14: 119-134.
HEFCE (2007)
Juznic P et al. (2010) Scientometric indicators: peer review, bibliometric methods and conflict of interest. *Scientometrics* 85:429-441.
Lamont M (2009) How Professors Think: inside the curious world of academic judgment. Harward Univ Press.
Ledesdorff L & Opthof T (2011) Remaining problems with the "New Crown Indicator" (MNCS) of the CWTS. Journal of Informetrics 5: 224-225.
Leydesdorff L & Opthof T (2010) Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. Journal of Informetrics 4: 423-430.
Marsh HW et al. (2008) Reliability, Validity, Bias and Generalizability. *American Psychologist* 63: 160-168.
Marsh, Bornmann, Mutz, Daniel, O'Mara (2009) Gender effects in the peer reviews of grant proposals: a comprehensive meta-analysis comparing traditional and multi-level approaches. *Review of Educational Research* 79:1290-1326
Sandström U & Hällsten M (2008) Persistent nepotism in peer review. Scientometrics 74:175-189,
Sandström U (2009) Cognitive bias in peer review: a new approach. Proceedings of the ISSI 2009, vol 2: 742-746.
Sandström U et al. (1997) Peers on peers: allocation policy and review procedures at theSwedish Research Council for Engineering Sciences. Stockholm.
Sandström, Wold, Jordansson (2010) *His Excellency: on funding of strong research environments* [In Swedish]. Delegation for Gender Equality in the University Sector. Report 2101:4. Stockholm.
Travis GDL & Collins HM (1991) New light on old boys: cognitive and institutional particularism in the peer-review system. Science Technology & Human Values 16: 322-341
Waltman L, van Eck NJ, van Leeuwen, TN, Visser MS, van Raan AFJ (2011) Towards a new crown indicator: some theoretical considerations. Journal of Informetrics 5:37-47.
Wennerås C & Wold A (1997) Nepotism and sexism in peer review. *Nature* 387: 341-343.
Wessely S (1998) Peer review of grant applications: what do we know? *The Lancet* 352: 301-305.