Chapter 7

The potential and problems of peer evaluation in higher education and research

Hans-Dieter Daniel*†[1], Sandra Mittag*† and Lutz Bornmann*
*ETH Zurich, Zaehringerstrasse 24, CH-8092 Zurich, Switzerland, and †University of Zurich, Evaluation Office, Muehlegasse 21, CH-8001 Zurich, Switzerland

Introduction

Science rests on peer review [1]. Peer review is the method by which grants are allocated, manuscripts published and study programmes improved. As 'gatekeepers' of science, the task of peers or colleagues asked to evaluate applications, manuscripts or study programmes is to ensure high standards in higher education and research. Peer review is regarded as the embodiment of the principle of mutual control [2]. Proponents of the system hold that peer review is more effective than any other known instrument for self-regulation in higher education and research. As stated by Eisenhart [3], "equals active in the same field are said to be in the best position to know whether quality standards have been met and a contribution to knowledge made". Thus the producers of science, the specialists, become the gatekeepers of science [4].

Peer evaluation in research means a process by which a selective jury of experts in a given scientific field is asked to evaluate the undertaking of scientific activity or its outcomes. Such a group of experts may be consulted as a group or individually, without the need for personal contacts among the evaluators [5]. Although there is evidence that peer review improves the quality of the reporting of research results [6,7], critics of peer review argue that: (i) reviewers rarely agree on whether or not to recommend that a manuscript be published or a grant be awarded, thus making for poor reliability of the peer-review process; (ii) recommendations and decisions in peer review are frequently biased, that is judgements are not based solely on scientific merit, but are also influenced by personal attributes of the authors, applicants or the reviewers themselves; (iii) the process lacks predictive validity, since there is little or no relationship between the judgements on applications or manuscripts and the subsequent usefulness of the proposed or published research to the scientific community, as indicated by the frequency of citations to the research in later scientific papers [8–15].

Usually, peer-review-based study programme evaluation begins with internal self-assessment, whereby an academic programme or institute conducts its own analysis of strengths and weaknesses for a self-evaluation report. The next stage is external evaluation. Here, peer reviewers conduct a site visit of the programmes or units under evaluation and prepare an external evaluation report. The follow-up stage entails implementation of the reviewers' recommendations [16]. Despite agreement on the general course of proceeding, national quality assurance systems differ greatly in the details [17,18]. The peer-review-based evaluation of study programmes has long come under critical fire. There are complaints, for instance in Germany (as well as in other countries), about the supposed high costs and burdens of evaluation, including both financial and personnel costs, and about the lack of consequences following evaluation [19]. In the face of scarce public funding, it is said that study programme evaluation serves merely to supply political decision-makers with information that is used for cost-cutting purposes and for changing the self-determination by professors into external control [20].

[1]email: daniel@evaluation.unizh.ch or daniel@gess.ethz.ch

In recent years, a number of published studies have addressed these and other criticisms that have been raised against peer evaluation in higher education and research. Studies that involved meta-evaluations of journal and grant peer-review procedures are presented below.

Meta-evaluation of journal and grant peer review

As an assessment tool, peer review is asked to be reliable, fair and valid [21]. Should a stock-taking of research on peer review find no evidence for the reliability, fairness and predictive validity of recommendations and decisions in peer review, one would have to question peer-reviewed contributions (publications and approved grant applications) as a measure of scientific advancement and of scientists' productivity.

Agreement among reviewers (reliability)

If a scientific contribution meets scientific standards and is a contribution to the advancement of science, it is reasonable to expect that two reviewers would agree on its value. Manuscripts and applications are rated reliably when there is a high level of agreement between independent reviewers. However, high agreement alone cannot result in high reliability, because a certain level of agreement can be expected to occur on the basis of chance alone.

Table 1 shows the results of some studies on reliability in the areas of journal (submission of manuscripts), meeting (submission of abstracts) and grant (submission of applications) peer review. The results indicate that the levels of inter-reviewer agreement, when corrected for chance, generally fall in the range 0.20–0.40. Coefficients between 0.20 and 0.40 indicate a relatively low level of reviewer agreement. At the same time, further examination of the data shows that agreement is substantially higher for recommendations for rejection than for acceptance. Reviewers are twice as likely to agree on rejection than on acceptance [22].

The following limitations of the studies on reviewer agreement have been pointed out [22]. First, experts have engaged in considerable debate on the subject of proper statistical tests for reviewer agreement studies, but the issues remain unresolved. Secondly, the analyses examined reviewers' final recommendations only. Reviewers' comments were usually not evaluated in detail. Therefore not much is known about the reasons behind reviewer disagreement. Thirdly, there is limited documentation on the need for or value of reviewer agreement.

Not all editors see reviewer disagreement as a negative factor; many see it as a positive method of evaluating a manuscript from a number of different perspectives. If reviewers are selected for their opposing viewpoints or expertise, a high degree of reviewer agreement should not be expected. It can even be argued that too much agreement is in fact a sign that the review process is not working well, that reviewers are not properly selected for diversity and that some are redundant. Whether the comments of reviewers are in fact based on different perspectives is a question that has been examined by only a few empirical studies [22]. One study, for example, showed that reviewers of the same manuscript simply commented on different aspects of the manuscript: "In the typical case, two reviews of the same paper had no critical point in common… [T]hey wrote about different topics, each making points that were appropriate and accurate. As a consequence, their recommendations about editorial decisions showed hardly any agreement" [23].

It has also been pointed out that the low degree of reviewer agreement reflects the low levels of cognitive consensus that exist at the research frontiers of all scientific disciplines [24]. At the frontiers of research, it is usually impossible to make an 'objective' evaluation of new work.

Fairness of the peer-review process

Journal submissions or grant applications are supposed to be judged solely on the basis of their scientific merit. The ideal Mertonian norm of universalism prescribes that the evaluation of new contributions should be based upon objective scientific criteria and not on the characteristics of the author, applicant or assessor [25]. Surveys of grant applicants and authors show, however, that they still have fears about a lack of objectivity in peer review. As many as 41% of the applicants of the NIHR (National Institute of Handicapped Research), now NIDRR

Table 1. Reliability: agreement among reviewers

Category	κ coefficient/intraclass correlation
Journals (submission of manuscripts)	
Social Problems [55]	0.40
Journal of Educational Psychology [56]	0.34
British Medical Journal [42]	0.31
American Sociological Review [57]	0.28
Physiological Zoology [57]	0.28
Journal of Personality and Social Psychology [58]	0.26
New England Journal of Medicine [59]	0.26
Law and Society Review [57]	0.17
Angewandte Chemie [60]	0.14
Physical Therapy [61]	0.12
Meetings (submission of abstracts)	
National Meeting of the American Association for the Study of Liver Disease [62]	0.24
Annual Meeting of the Orthopaedic Trauma Association [63]	0.23
Research funding institutions (submission of applications)	
American Heart Association [64]	0.37
National Science Foundation (solid state physics) [65]	0.32
Heart and Stroke Foundation, Medical Research Council of Canada [66]	0.29

(National Institute on Disability and Rehabilitation Research) (Washington, DC, U.S.A.), disagreed with the statement "the peer reviewer comments were fair" [26]. In a survey of article authors of the professional journals *Academy of Management Journal* and *Academy of Management Review*, about one out of ten survey respondents was dissatisfied with the reviewers' objectivity [27].

Reviews of peer review research [9,28–32] name up to 25 different potential sources of bias in peer review. In these studies, it is usual to call any feature of an assessor's cognitive or attitudinal mindset that could interfere with an objective judgment a bias [33]. Factors that appear to bias assessors' objective judgements with respect to a manuscript or an application include nationality, gender of the author or applicant and the area of research from which the work originates. Other studies show that replication studies and research that leads to statistically insignificant findings stand a rather low chance of being judged favourably by peer reviewers. Research on bias in peer review faces two serious problems. First, the research findings on bias are inconsistent. For example, some studies investigating gender bias in journal review processes point out that women scientists are at a disadvantage (see Table 2). However, a similar number of studies report no gender effects or mixed results.

Secondly, it is almost impossible to establish unambiguously whether work from a particular group of scientists (e.g. junior or senior scientists) receives better reviews and thus a higher acceptance rate due to preferential biases affecting the review and decision-making process, or if favourable review and favourable judgements in peer review are a simple consequence of the high scientific quality of the corresponding manuscripts or applications.

Presumably, it will never be possible to eliminate all doubts regarding the fairness of the reviewing process. Because reviewers are humans, their behaviour, whether performing their salaried duties, enjoying their leisure time or writing reviews, is influenced by factors that cannot be predicted, controlled or standardized [34]. Therefore it is important that the process of peer reviewing should be studied continuously. Any evidence of bias in judgements should be uncovered for purposes of correction and modification of the process [35,36].

Table 2. Results of empirical studies on gender effects in journal peer review

Role in question	Study results indicate a gender effect	Study results indicate no gender effect	Study results on gender effect are mixed
Author	Petty and Fleming, 1999 [67]	Caelleigh et al., 2003 [68]	Tregenza, 2002 [69]
	Lloyd, 1990 [70]	Gilbert et al., 1994 [71]	Levenson et al., 1975 [72]
	Paludi and Strayer, 1985 [73]	Blank, 1991 [74]	
	Paludi and Bauer, 1983 [75]	Patterson et al., 1987 [76]	
	Sahner, 1982 [77]	Bernhard, 1980 [78]	
	Ward, 1981 [79]		
	Ferber and Teiman, 1980 [80]		
	Goldberg, 1968 [81]		
Reviewer	Blank, 1991 [74]	del Carmen Davo et al., 2003 [82]	
	Lloyd, 1990 [70]	Gilbert et al., 1994 [71]	
	Paludi and Strayer, 1985 [73]	Nylenna et al., 1994 [83]	
	Paludi and Bauer, 1983 [75]		
	Ward, 1981 [79]		
	Levenson et al., 1975 [72]		

Predictive validity of the peer-review process

What is peer review for? One answer to this question is that it is a method for selecting the best grant applications for funding and the best manuscripts for journal publication. It is difficult to test this aim, because there is no agreed upon definition of what constitutes a good manuscript or a good application [37]. A conventional approach is to use citation counts as a proxy for research contributions, since they measure the international impact of the research by individuals or groups of scientists [38–40]. Critics of peer review [9,15] claimed that manuscript refereeing is without validity in forecasting the subsequent usefulness of a work to scientists as reflected in citations of the work in other scientific papers. On the contrary, based on citation rates for accepted manuscripts and previously rejected manuscripts published elsewhere, editorial decisions in all of the four existing studies [41–44] showed a high degree of predictive validity. Similar results were reported for grant peer review [45–48]. The most serious limitation of research on the predictive validity of recommendations and decisions in peer-review processes is the very small number of studies. In the following, we present as examples two studies on the predictive validity of peer review.

In a study on the peer-review process of *Angewandte Chemie* [43], one of the most important chemistry journals worldwide, we traced the fate of rejected manuscripts. Of the manuscripts rejected by *Angewandte Chemie*, 71% were later published in other journals. These 'rejects' were published in a total of 39 journals. However, none of the rejected manuscripts appeared in a journal that had a greater impact factor than *Angewandte Chemie*. In addition, papers accepted by *Angewandte Chemie* were on average cited twice as frequently as manuscripts that had been rejected on the basis of reviewers' recommendations, but were later published elsewhere. The editor of the *Journal of Clinical Investigation*, Jean Wilson, the editor of the *British Medical Journal*, Stephen Lock, and the editor of *Cardiovascular Research*, Tobias Opthof, established very similar results [41,42,44].

A review of the study on the peer-review process of *Angewandte Chemie* [49] described above raised an argument against this form of validity test, stating that papers accepted by *Angewandte Chemie* may have been cited on average more frequently than those published elsewhere simply because they appeared in a journal with a high JCR (Journal Citation Report) Impact Factor (provided by Thomson Scientific, Philadelphia, PA, U.S.A.), i.e. in a journal with high visibility. However, the citation rates of journal articles do not seem to be detectably influenced by the status of the journals in which they are published [50].

In a study on committee peer review for the selection of doctoral (Ph.D.) and postdoctoral research fellowship recipients, we analysed the predictive validity of the peer-review process followed by the B.I.F. (Boehringer Ingelheim Fonds, Heidesheim, Germany), a foundation for the promotion of basic research in biomedicine, in two steps [45]. In a first step, we examined 2039 articles by 120 former fellowship recipients that had been published between the date of approval of the fellowship and December 2000. The results showed that the research articles by B.I.F. fellows were cited considerably more often than the 'average' paper (average citation rate) published in the journal sets corresponding to the fields 'Multidisciplinary', 'Molecular Biology and Genetics' and 'Biology and Biochemistry' in ESI (Essential Science Indicators) from Thomson Scientific (most of the fellows publish within these fields).

In the second step, we conducted a citation analysis for articles published *previous* to the applicants' approval or rejection for a B.I.F. fellowship. On the basis of the model estimation (negative binomial regression model), journal articles that had been published by applicants approved for a fellowship award (64 applicants) prior to applying for the B.I.F. fellowship award can be expected to have 37% (straight counts of citations) and 49% (complete counts of citations) more citations than articles that had been published by rejected applicants (333 applicants). Furthermore, comparison with international scientific reference values revealed that (i) articles published by successful and non-successful applicants were cited considerably more often than the 'average' publication, and (ii) excellent research performance can be expected more from successful than non-successful applicants. All in all, the findings of the two steps of the comprehensive analyses

confirmed that the foundation is not only achieving its goal of selecting the best junior scientists for fellowship awards, but also successfully attracting highly talented young scientists to apply for B.I.F. fellowships.

Meta-evaluation of peer-review-based study programme evaluation

The International Centre for Higher Education Research Kassel (INCHER-Kassel, Germany) initiated an analysis of procedures and the effectiveness of the evaluation processes of the Central Evaluation and Accreditation Agency Hanover (ZEvA, Zentrale Evaluations- und Akkreditierungsagentur Hannover) and the Consortium of Universities in Northern Germany (VNU, Verbund Norddeutscher Universitäten) in the evaluation of study programmes. The study is the first comprehensive and representative investigation of peer-review-based study programme evaluations in Germany, and it assesses the two most tried-and-tested and best-known evaluation procedures in Germany [in all, there are eight evaluation agencies in Germany for systematic study programme evaluation at HEIs (higher education institutions)] from multiple perspectives and using multiple methods [51]. VNU is a consortium of six German universities, and ZEvA is a common agency for Lower Saxony HEIs; both have conducted peer-review-based study programme evaluations since the mid-1990s. They completed the first evaluation cycle in 2001.

The study data were collected in a questionnaire survey by mail of all former external reviewers and members of institutes ('members of institutes' refers to all members of an evaluation working group formed within an institute under evaluation, including students) who participated in evaluations conducted by ZEvA and VNU. A total of 648 returned questionnaires could be included in the analysis, with a response rate of 41%. In addition to the questionnaire survey, 33 interviews were conducted and examined using content analysis. The interview participants were university heads and the authorized agents or contact partners at the different HEIs, the spokesperson of VNU, the scientific director at ZEvA and the managing directors and staff members of VNU and ZEvA.

Overall assessment of the evaluation procedures

To tap overall assessments, the members of institutes and reviewers were asked whether the peer-review-based study programme evaluation (internal evaluation, external evaluation and implementation of recommendations) proved to be useful and effective, whether the evaluation process achieved the goals of quality assurance and improvement, whether the results of the evaluation were commensurate to the effort required, whether they were satisfied overall with the process and whether participation in the evaluation process proved worthwhile for them personally.

As shown in Figure 1, the institute members (83%) and reviewers (96%) stated that the evaluations of study programmes proved to be useful and effective. The majority of the respondents were satisfied overall with the evaluations conducted at the individual universities (68% of institute members and 95% of reviewers). The majority of the respondents also found that the evaluation process achieved the goals of quality assurance and improvement (65% of institute members and 93% of reviewers) and that participation in the evaluation proved personally rewarding (61% of institute members and 93% of reviewers). Whereas 82% of the reviewers saw the effort entailed as commensurate to the results of the evaluation, the majority of the institute members (55%) found the effort required to be disproportionate to the results. All in all, however, as the results show, peer-review-based study programme evaluation proved to be useful and effective in the opinion of the reviewers and the members of the institutes.

As an alternative to the multi-stage procedure, some German universities conduct evaluations of study programmes in only one stage. The one-stage procedure can be either an external evaluation or an internal evaluation only. In the case of an internal evaluation only, a university consultant may be asked to provide professional support. Exclusively external evaluations include, among others, structural evaluations by management consultants [52].

The reviewers and members of institutes were asked to indicate possible preferences for

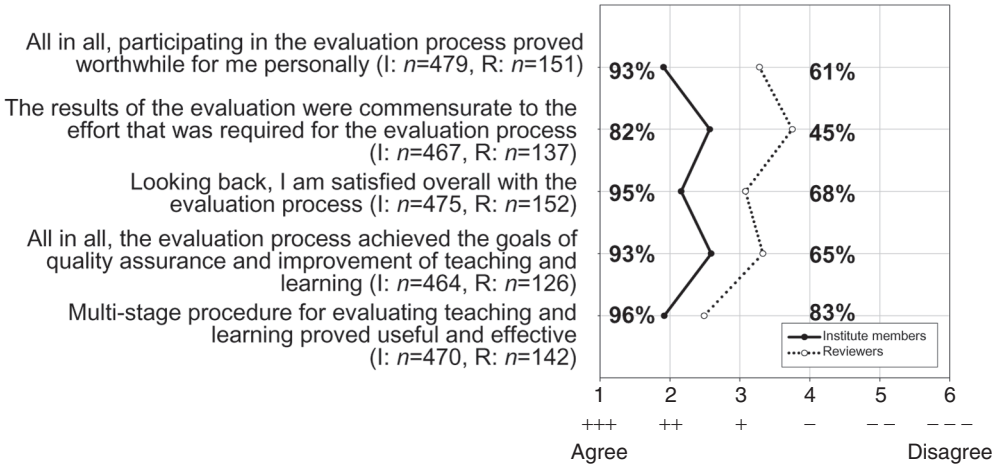

Figure 1. Overall assessment of the evaluation process by institute members (I) and reviewers (R) (arithmetic average and relative frequencies; the values show the percentage of institute members and reviewers that checked response 1, 2 or 3 on the questionnaire item; the assessments are sorted in descending order according to mean differences between institute members and reviewers).

certain types of one-stage procedures over the multi-stage procedure. As Table 3 shows, 12% of the respondents preferred to see study programme evaluation conducted in the form of an internal evaluation exclusively and 24% in the form of an internal evaluation with the support of a university consultant (for this question, multiple answers were possible). As an alternative to a purely internal evaluation, 22% preferred to see study programme evaluation conducted in the form of an external evaluation exclusively, by review committee, and 5% in the form of an external evaluation exclusively, by one reviewer.

The interview participants also rated the evaluation procedure overall as positive. The majority of the interviewees believed that the evaluations were necessary and that they proved to be useful and effective. They emphasized as particular strengths of the evaluations that in their design the evaluations were university-independent, self-critical and well-structured and organized. On the other hand, some interview participants offered the criticism that the purpose of the evaluation was clarified insufficiently, that the personnel situation in the institute was not taken into account sufficiently and that the evaluation did not take a sufficiently international orientation.

Linking study programme evaluations with rankings, university funding, evaluation of research and accreditation

In our meta-evaluation of peer-review-based study programme evaluation, the written questionnaire respondents were also asked (i) whether there should be a ranking of the universities participating in the evaluation, (ii) whether they would like to see evaluation results linked to funding decisions, and whether the study programme evaluation should be linked to (iii) the evaluation of research or (iv) accreditation.

Evaluation of study programmes and the issue of ranking

In line with common practice in European higher education [16], the evaluation procedures of ZEvA and VNU do not aim to produce a ranking of the participating institutions. An item on the questionnaire asked whether they had nonetheless feared during the different phases of the evaluation that their institute would be ranked. The findings show that 62% of institute members and 80% of reviewers had not gained the impression that the evaluation was aiming towards ranking.

Table 3. Percentage of evaluation participants that would prefer one-stage evaluation procedures to multi-stage evaluation procedures (multiple answers possible; in absolute and relative frequencies)

Type of one-stage evaluation procedure	Frequency	
	Absolute	%
Evaluation of teaching and research study programmes: internal evaluation, with support by university consultant (n=538)	128	24
Evaluation of teaching and research study programmes : external evaluation only, by review committee (n=554)	121	22
Evaluation of teaching and research study programmes: internal evaluation only (n=524)	62	12
Evaluation of teaching and research study programmes: external evaluation only, by one reviewer (n=523)	26	5

Both the questionnaire respondents and interview participants were asked whether they would have found a ranking of the participating HEIs desirable. The clear majority of interview participants spoke against ranking. Nearly three-quarters (72%) of the survey respondents rejected ranking. Among the 28% that found ranking desirable, members of institutes (31%) were somewhat more strongly represented than reviewers (19%). Overall, the findings as to ranking of HEIs participating in evaluation make it clear that the majority of people involved in evaluations neither fear nor are in favour of ranking.

Evaluation of study programmes and the issue of linkage to funding

In the discussion on higher education policies in Germany, some voices have considered linking the amount of funding allocated to an institute to the results of evaluations. This linkage could result in funding increases or cuts. On the survey questionnaire, the respondents were asked to assess these proposals. They were asked whether evaluation results should be linked to funding increases only, to cuts in funding only, to both increases and reductions in funding, or not linked to funding at all (see Table 4). Regarding the linkage issue, members of institutes (42%) voted somewhat more frequently than reviewers (33%) for the alternative, namely no linkage to funding, while reviewers (46%) voted somewhat more frequently than the institute members (37%) in favour of linking evaluation results to funding increases or reductions.

Approximately 20% of both groups favoured linkage of evaluation results to funding increases only, and only 1% (institute members) and 2% (reviewers) approved of linkage to cuts in funding. The majority of the interview participants saw evaluation results becoming linked to funding decisions in future. Most of these interview participants viewed both positive and negative sanctions as appropriate. Still, a considerable number of the interview participants thought that linkage of evaluation to funding decisions would not be appropriate.

Evaluation of study programmes and the issue of linkage to evaluation of research

ZEvA and VNU evaluate study programmes. We asked participants of both ZEvA and VNU evaluations whether the evaluation of study programmes and the evaluation of research should in future be conducted jointly or, as has been the case up to now, in separate evaluations. More than half of the questionnaire respondents (55%) indicated a preference for joint evaluation of teaching and research, while 39% voted for separate evaluations of these areas (6% of the respondents had no opinion). On this issue, the institute members and the reviewers are in agreement.

The majority of the interview participants, in contrast, rejected the idea of a joint evaluation procedure for the two areas. The main reasons given by the interview participants were the following: (i) teaching would decrease in importance in comparison with research; (ii) research has a stronger interdisciplinary orientation; (iii) evaluations of research focus more strongly on the performance of

Table 4. Issue of linking results of evaluation to funding of institutions or programmes, by institute members and reviewers (in absolute and relative frequencies; the assessments are sorted in descending order according to percentages among institute members)

	Institute members		Reviewers	
Linking evaluation to funding	**Absolute**	**%**	**Absolute**	**%**
Evaluation results should never be linked to funding	193	42	47	33
Evaluation results should be linked to increases and reductions in funding	169	37	65	46
Evaluation results should be linked to increases in funding only	95	20	27	19
Evaluation results should be linked to reductions in funding only	3	1	3	2
Total	460	100	142	100

individuals and less on the department as a whole; (iv) the objectives of the two evaluation procedures are too different; and (v) a joint evaluation procedure would require too much effort.

The interview participants found, however, that dovetailing study programme evaluation with research evaluation made sense. They stated that the time points of conducting the two evaluations should be co-ordinated and that a common statistical database should be used. For future decision-making affecting an entire department, co-ordination would make it easier for the results of both evaluations to be considered.

Evaluation of study programmes and the issue of linkage to accreditation

The interview participants were rather reserved in their judgements on linking study programme evaluations to accreditation. The majority of the interview participants stated that the two procedures target different objectives and that the relationship between study programme evaluation and accreditation would first have to be clarified. But considering the effort associated with both accreditation procedures and study programme evaluation, the interview participants were in favour of at least some kind of connection (for example, through the use of the same documents). A proportion of the interview participants stated the opinion that, in future, study programme evaluation, accreditation and evaluation of research should be conducted jointly or in co-ordination in order to minimize the effort and expense required for different procedures conducted in parallel.

Discussion

In recent years, a number of published studies have taken up and investigated the criticisms that have been raised against peer evaluation in higher education and research. Some important studies were presented in the sections above.

Research on the predictive validity of peer evaluation in research indicates that peer review is generally a credible method for *ex ante* evaluation of manuscripts and grant applications. But our overview of the reliability and fairness of peer review shows that there are also problems with peer review. However, despite its flaws, having scientists judge each other's work is widely considered to be the 'least bad way' to weed out weak manuscripts or research proposals and improve promising ones [53]. Therefore *ex ante* peer review should be used for the evaluation of manuscripts and grant applications and should be supplemented *ex post* with bibliometrics and other metrics of science to yield a broader and powerful methodology for assessment of scientific advancement [35,54].

The meta-evaluation of peer-review-based study programme evaluation demonstrated that former participants (reviewers and those reviewed) in the ZEvA and VNU evaluations were satisfied all in all with the evaluations and believed that the goals of quality assurance and improvement had been achieved. Somewhat over half of the members of institutes that were surveyed by questionnaire, however, found that the results of the evaluations did

not justify the heavy work burden that the process entailed. In other words, a substantial proportion of institute members had concerns about the cost-benefit value of the evaluation process.

To sum up, contrary to the diverse criticisms that are still being raised against peer evaluations in higher education and research, the findings of the meta-evaluation studies confirm that peer review enjoys wide acceptance and can be seen as useful.

References

1 Ziman, J. (2000) *Real Science: What it is, and What it Means*, Cambridge University Press, Cambridge
2 Polanyi, M. (1966) *The Tacit Dimension*, Doubleday, New York
3 Eisenhart, M. (2002) The paradox of peer review: admitting too much or allowing too little? *Research in Science Education* **32**(2), 241–255
4 McClellan, J.E. (2003) Specialist control: the publications committee of the Académie Royal des Sciences (Paris) 1700–1793. *Transactions of the American Philosophical Society* **93**(3)
5 Geisler, E. (2000) *The Metrics of Science and Technology*, Quorum Books, Westport
6 Goodman, S.N., Berlin, J., Fletcher, S.W. and Fletcher, R.H. (1994) Manuscript quality before and after peer review and editing at *Annals of Internal Medicine*. *Annals of Internal Medicine* **121**(1), 11–21
7 Pierie, J.P.E.N., Walvoort, H.C. and Overbeke, A.J.P.M. (1996) Readers' evaluation of effect of peer review and editing on quality of articles in the *Nederlands Tijdschrift voor Geneeskunde*. *Lancet* **348**(9040), 1480–1483
8 Finn, C.E. (2002) The limits of peer review. *Education Week* **21**(34), 30–34
9 Ross, P.F. (1980) *The Sciences' Self-Management: Manuscript Refereeing, Peer Review, and Goals in Science*, The Ross Company, Todd Pond
10 Abate, T. (1995) What's the verdict on peer review? *Ethics Research* **1**(1), 1
11 Roy, R. (1985) Funding science: the real defects of peer-review and an alternative to it. *Science, Technology, and Human Values* **10**(3), 73–81
12 Moran, G. (1998) *Silencing Scientists and Scholars in Other Fields: Power, Paradigm Controls, Peer Review, and Scholarly Communication*, Ablex Publishing Corporation, London
13 Langfeldt, L. (2004) Expert panels evaluating research: decision-making and sources of bias. *Research Evaluation* **13**(1), 51–62
14 Horrobin, D.F. (2001) Something rotten at the core of science? *Trends Pharmacological Science* **22**(2), 51–52
15 Eysenck, H.J. and Eysenck, S.B.G. (1992) Peer review: advice to referees and contributors. *Personality and Individual Differences* **13**(4), 393–399
16 Danish Evaluation Institute (2003) *Quality Procedures in European Higher Education: an ENQA Survey*, European Network for Quality Assurance in Higher Education, Helsinki
17 Brennan, J. and Shah, T. (2000) *Managing Quality in Higher Education: an International Perspective on Institutional Assessment and Change*, Organisation for Economic Cooperation and Development (OECD), Buckingham
18 Billing, G. (2004) International comparisons and trends in external quality assurance of higher education: commonality or diversity? *Higher Education* **47**(1), 113–137
19 Berthold, C. (2002) Von der Evaluation zur strategischen Hochschulentwicklung – 16 Thesen. In *Qualitätssicherung an Hochschulen: Theorie und Praxis* (Reil, T. and Winter, M., eds), pp. 160–165, Bertelsmann, Bielefeld
20 Erche, B. (2003) Evaluation als politisches Steuerungsinstrument. In *Evaluation der Universitären Lehre in der Medizin: Gegenstände, Methoden, Konsequenzen* (Neuser, J. and Urban, R., eds), pp. 3–7, Shaker, Aachen
21 Hackett, E.J. and Chubin, D.E. (2003) Peer review for the 21st century: applications to education research. Paper presented at a National Research Council Workshop on *Peer Review of Education Research Grant Applications: Implications, Considerations, and Future Directions*, February 25–26 Washington, DC (http://www7.nationalacademies.org/core/Peer%20Review.html)
22 Weller, A.C. (2002) *Editorial Peer Review: its Strengths and Weaknesses*, Information Today, Medford
23 Fiske, D.W. and Fogg, L. (1990) But the reviewers are making different criticisms of my paper: diversity and uniqueness in reviewer comments. *American Psychologist* **45**(5), 591–598
24 Cole, J.R. (2000) The role of journals in the growth of scientific knowledge. In *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield* (Cronin, B. and Atkins, H.B., eds), pp. 109–142, Information Today, Medford
25 Merton, R.K. (1942) Science and technology in a democratic order. *Journal of Legal and Political Sociology* **1**(1–2), 115–126
26 Fuhrer, M.J. and Grabois, M. (1985) Grant application and review procedures of the National Institute of Handicapped Research: survey of applicant and peer reviewer opinions. *Archives of Physical Medicine and Rehabilitation* **66**(5), 318–321
27 Bedeian, A.G. (2003) The manuscript review process: the proper roles of authors, referees, and editors. *Journal of Management Inquiry* **12**(4), 331–338
28 Sharp, D.W. (1990) What can and should be done to reduce publication bias: the perspective of an editor. *JAMA: the Journal of the American Medical Association* **263**(10), 1390–1391
29 Owen, R. (1982) Reader bias. *JAMA: the Journal of the American Medical Association* **247**(18), 2533–2534
30 Hojat, M., Gonnella, J.S. and Caelleigh, A.S. (2003) Impartial judgment by the "gatekeepers" of science: fallibility and accountability in the peer review process. *Advances in Health Sciences Education* **8**(1), 75–96
31 Pruthi, S., Jain, A., Wahid, A., Mehra, K. and Nabi, S.A. (1997) Scientific community and peer review system: a

case study of a central government funding scheme in India. *Journal of Scientific and Industrial Research* **56**(7), 398–407
32 Wood, F.Q. and Wessely, S. (2003) Peer review of grant applications: a systematic review. In *Peer Review in Health Sciences* (Godlee, F. and Jefferson, T., eds), pp. 14–44, BMJ Publishing Group, London
33 Shatz, D. (2004) *Peer Review: a Critical Inquiry*, Rowman and Littlefield, Lanham
34 Shashok, K. (2005) Standardization vs diversity: how can we push peer review research forward? *Medscape General Medicine* **7**(1), 11
35 Geisler, E. (2001) The mires of research evaluation. *Scientist* **15**(10), 39
36 Godlee, F. and Dickersin, K. (2003) Bias, subjectivity, chance, and conflict of interest. In *Peer Review in Health Sciences* (Godlee, F. and Jefferson, J., eds), pp. 91–117, BMJ Publishing Group, London
37 Smith, R. (2006) Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine* **99**(4), 178–182
38 Cole, J.R. (2000) A short history of the use of citations as a measure of the impact of scientific and scholarly work. In *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield* (Cronin, B. and Atkins, H.B., eds), pp. 281–300, Information Today, Medford
39 van Raan, A.F.J. (2004) Measuring science: capita selecta of current main issues. In *Handbook of Quantitative Science and Technology Research: the Use of Publication and Patent Statistics in Studies of S&T Systems* (Moed, H.F., Glänzel, W. and Schmoch, U., eds), pp. 19–50, Kluwer Academic Publishers, Dordrecht
40 Daniel, H.-D. (2005) Publications as a measure of scientific advancement and of scientists' productivity, *Learned Publishing* **18**(2), 143–148
41 Wilson, J.D. (1978) Peer review and publication. *Journal of Clinical Investigation* **61**(6), 1697–1701
42 Lock, S. (1985) *A Difficult Balance: Editorial Peer Review in Medicine*, ISI Press, Philadelphia
43 Daniel, H.-D. (1993) *Guardians of Science: Fairness and Reliability of Peer Review*, Wiley-VCH, Weinheim
44 Opthof, T., Furstner, F., van Geer, M. and Coronel, R. (2000) Regrets or no regrets? No regrets! The fate of rejected manuscripts. *Cardiovascular Research* **45**(1), 255–258
45 Bornmann, L. and Daniel, H.-D. (2005) Selection of research fellowship recipients by committee peer review: analysis of reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics* **63**(2), 297–320
46 Bornmann, L. and Daniel, H.-D. (2005) Committee peer review at an international research foundation: predictive validity and fairness of selection decisions on post-graduate fellowship applications. *Research Evaluation* **14**(1), 15–20
47 Mavis, B. and Katz, M. (2003) Evaluation of a program supporting scholarly productivity for new investigators. *Academic Medicine* **78**(7), 757–765
48 Bornmann, L. and Daniel, H.-D. (2006) Selecting scientific excellence through committee peer review: a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics* **68**(3), 427–440
49 Cicchetti, D.V. (1999) Guardians of science: fairness and reliability of peer review. *Journal of Clinical and Experimental Neuropsychology* **21**(3), 412–421
50 Seglen, P.O. (1994) Causal relationship between article citedness and journal impact. *Journal of the American Society Information Science and Technology* **45**(1), 1–11
51 Bornmann, L., Mittag, S. and Daniel, H.-D. (2006) Quality assurance in higher education: meta-evaluation of multi-stage evaluation procedures in Germany. *Higher Education* **52**(4), 687–709
52 Webler, W.-D. (1998) Das Bielefelder Modell zur Evaluation der Lehre als Organisationsberatung durch Hochschulforscher. In *Evaluation und Qualitätssicherung an den Hochschulen in Deutschland: Stand und Perspektiven: Nationales Expertenseminar der Hochschulrektorenkonferenz, Bonn, 29 May 1998*, pp. 189–196, Hoc hschulrektorenkonferenz, Bonn
53 Enserink, M. (2001) Peer review and quality: a dubious connection? *Science* **293**(5538), 2187–2188
54 van Raan, A.F.J. (1996) Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics* **36**(3), 397–420
55 Smigel, E.O. and Ross, H.L. (1970) Factors in editorial decision. *American Sociologist* **5**(1), 19–21
56 Marsh, H.W. and Ball, S. (1981) Interjudgmental reliability of reviews for the *Journal of Educational Psychology*. *Journal of Educational Psychology* **73**, 872–880
57 Hargens, L.L. and Herting, J.R. (1990) A new approach to referees' assessments of manuscripts. *Social Science Research* **19**(1), 1–16
58 Scott, W.A. (1974) Interreferee agreement on some characteristics of manuscripts submitted to journal of personality and social psychology. *American Psychologist* **29**(9), 698–702
59 Ingelfinger, F.J. (1974) Peer review in biomedical publication. *American Journal of Medicine* **56**(5), 686–692
60 Daniel, H.-D. (1993) An evaluation of the peer-review process at Angewandte Chemie. *Angewandte Chemie International Edition* **32**(2), 234–238
61 Bohannon, R.W. (1986) Agreement among reviewers. *Physical Therapy* **66**(9), 1431–1432
62 Cicchetti, D.V. and Conn, H.O. (1976) A statistical analysis of reviewer agreement and bias in evaluating medical abstracts. *The Yale Journal of Biology and Medicine* **49**(4), 373–383
63 Bhandari, M., Templeman, D. and Tornetta, P. (2004) Interrater reliability in grading abstracts for the Orthopaedic Trauma Association. *Clinical Orthopaedics and Related Research* **423**, 217–221
64 Wiener, S., Urivetsky, M., Bregman, D., Cohen, J., Eich, R., Gootman, N., Gulotta, S., Taylor, B., Tuttle, R., Webb, W. and Wright, J. (1977) Peer review: inter-reviewer agreement during evaluation of research grant evaluations. *Clinical Research* **25**(5), 306–311
65 Cicchetti, D.V. (1991) The reliability of peer review for manuscript and grant submissions: a cross-disciplinary investigation. *Behavioural and Brain Science* **14**(1), 119–135
66 Hodgson, C. (1997) How reliable is peer review? An examination of operating grant proposals simultaneously

submitted to two similar peer review systems. *Journal of Clinical Epidemiology* **50**(11), 1189–1195
67 Petty, R.E. and Fleming, M.A. (1999) The review process at *PSPB*: correlates of interreviewer agreement and manuscript acceptance. *Personality and Social Psychology Bulletin* **25**(2), 188–203
68 Caelleigh, A.S., Hojat, M., Steinecke, A. and Gonnella, J.S. (2003) Effects of reviewers' gender on assessments of a gender-related standardized manuscript. *Teaching and Learning in Medicine* **15**(3), 163–167
69 Tregenza, T. (2002) Gender bias in the refereeing process? *Trends in Ecology and Evolution* **17**(8), 349–350
70 Lloyd, M.E. (1990) Gender factors in reviewer recommendations for manuscript publication. *Journal of Applied Behavioural Analysis* **23**(4), 539–543
71 Gilbert, J.R., Williams, E.S. and Lundberg, G.D. (1994) Is there gender bias in *JAMA*'s peer review process. *JAMA: the Journal of the American Medical Association* **272**(2), 139–142
72 Levenson, H., Burford, B., Bonno, B. and Davis, L. (1975) Are women still prejudiced against women? A replication and extension of Goldberg's Study. *Journal of Psychology* **89**(1), 67–71
73 Paludi, M.A. and Strayer, L.A. (1985) What's in an authors name? Differential evaluations of performance as a function of author's name. *Sex Roles* **12**(3–4), 353–361
74 Blank, R.M. (1991) The effects of double-blind versus single-blind reviewing: experimental evidence from the *American Economic Review*. *American Economic Review* **81**(5), 1041–1067
75 Paludi, M.A. and Bauer, W.D. (1983) Goldberg revisited: what's in an authors name. *Sex Roles* **9**(3), 387–390
76 Patterson, S.C., Bailey, M.S., Martinez, V.J. and Angel, S.C. (1987) Report of the managing editor of the *American Political Science Review*, 1986–1987. *PS* **20**(4), 1006–1016
77 Sahner, H. (1982) Zur Selektivität von Herausgebern: eine Input-output-Analyse der *Zeitschrift für Soziologie*. *Zeitschrift für Soziologie* **11**(1), 82–98
78 Bernard, H.R. (1980) Report from the editor. *Human Organization* **39**(4), 366–369
79 Ward, C. (1981) Prejudice against women: who, when, and why? *Sex Roles* **7**(2), 163–171
80 Ferber, M.A. and Teiman, M. (1980) Are women economists at a disadvantage in publishing journal articles? *Eastern Economic Journal* **6**(3–4), 189–193
81 Goldberg, P. (1968) Are women prejudiced against women? *Transaction* **5**(1), 28–30
82 del Carmen Davo, M., Vives, C. and Álvarez-Dardet, C. (2003) Why are women underused in the *JECH* peer review process? *Journal of Epidemiology and Community Health* **57**(12), 936-937
83 Nylenna, M., Riis, P. and Karlsson, Y. (1994) Multiple blinded reviews of the same two manuscripts: effects of referee characteristics and publication language. *JAMA: the Journal of the American Medical Association* **272**(2), 149–151