Chapter 7

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

Hans-Dieter Daniel*†, 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
The potential and problems of peer evaluation

(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>
<thead>
<tr>
<th>Table 1. Reliability: agreement among reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category</td>
</tr>
<tr>
<td>Journals (submission of manuscripts)</td>
</tr>
<tr>
<td><em>Social Problems</em> [55]</td>
</tr>
<tr>
<td><em>Journal of Educational Psychology</em> [56]</td>
</tr>
<tr>
<td><em>British Medical Journal</em> [42]</td>
</tr>
<tr>
<td><em>American Sociological Review</em> [57]</td>
</tr>
<tr>
<td><em>Physiological Zoology</em> [57]</td>
</tr>
<tr>
<td><em>Journal of Personality and Social Psychology</em> [58]</td>
</tr>
<tr>
<td><em>New England Journal of Medicine</em> [59]</td>
</tr>
<tr>
<td><em>Law and Society Review</em> [57]</td>
</tr>
<tr>
<td><em>Angewandte Chemie</em> [60]</td>
</tr>
<tr>
<td><em>Physical Therapy</em> [61]</td>
</tr>
<tr>
<td>Meetings (submission of abstracts)</td>
</tr>
<tr>
<td>National Meeting of the American Association for the Study of Liver Disease [62]</td>
</tr>
<tr>
<td>Annual Meeting of the Orthopaedic Trauma Association [63]</td>
</tr>
<tr>
<td>Research funding institutions (submission of applications)</td>
</tr>
<tr>
<td>American Heart Association [64]</td>
</tr>
<tr>
<td>National Science Foundation (solid state physics) [65]</td>
</tr>
<tr>
<td>Heart and Stroke Foundation, Medical Research Council of Canada [66]</td>
</tr>
</tbody>
</table>

(© The Authors Volume compilation © 2007 Portland Press Ltd)
### Table 2. Results of empirical studies on gender effects in journal peer review

<table>
<thead>
<tr>
<th>Role in question</th>
<th>Study results indicate a gender effect</th>
<th>Study results indicate no gender effect</th>
<th>Study results on gender effect are mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paludi and Strayer, 1985 [73]</td>
<td>Blank, 1991 [74]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paludi and Bauer, 1983 [75]</td>
<td>Patterson et al., 1987 [76]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sahner, 1982 [77]</td>
<td>Bernhard, 1980 [78]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ward, 1981 [79]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ferber and Teiman, 1980 [80]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goldberg, 1968 [81]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reviewer</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blank, 1991 [74]</td>
<td>del Carmen Davo et al., 2003 [82]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lloyd, 1990 [70]</td>
<td>Gilbert et al., 1994 [71]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paludi and Strayer, 1985 [73]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paludi and Bauer, 1983 [75]</td>
<td>Nylenna et al., 1994 [83]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ward, 1981 [79]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levenson et al., 1975 [72]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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
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.
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>
<thead>
<tr>
<th>Type of one-stage evaluation procedure</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation of teaching and research study programmes: internal evaluation, with support by university consultant (n=538)</td>
<td>128 24</td>
</tr>
<tr>
<td>Evaluation of teaching and research study programmes: external evaluation only, by review committee (n=554)</td>
<td>121 22</td>
</tr>
<tr>
<td>Evaluation of teaching and research study programmes: internal evaluation only (n=524)</td>
<td>62 12</td>
</tr>
<tr>
<td>Evaluation of teaching and research study programmes: external evaluation only, by one reviewer (n=523)</td>
<td>26 5</td>
</tr>
</tbody>
</table>
The potential and problems of peer evaluation

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)

<table>
<thead>
<tr>
<th>Linking evaluation to funding</th>
<th>Institute members</th>
<th>Reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Absolute</td>
<td>%</td>
</tr>
<tr>
<td>Evaluation results should never be linked to funding</td>
<td>193</td>
<td>42</td>
</tr>
<tr>
<td>Evaluation results should be linked to increases and reductions in funding</td>
<td>169</td>
<td>37</td>
</tr>
<tr>
<td>Evaluation results should be linked to increases in funding only</td>
<td>95</td>
<td>20</td>
</tr>
<tr>
<td>Evaluation results should be linked to reductions in funding only</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>460</td>
<td>100</td>
</tr>
</tbody>
</table>

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

A Difficult Balance: Editorial Peer Review in Medicine, ISI Press, Philadelphia
Bormann, 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
Shashok, K. (2005) Standardization vs diversity: how can we push peer review research forward? Medscape General Medicine 7(1), 11
Bormann, 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
submitted to two similar peer review systems. *Journal of Clinical Epidemiology* 50(11), 1189–1195


