

Reliability, Fairness, and Predictive Validity of the Peer Review Process for the Selection of Research Fellowship Recipients of the Boehringer Ingelheim Fonds

Lutz Bornmann and Hans-Dieter Daniel

Introduction

Peer review is the principal mechanism for quality control in funding of academic science with increasing usage through the general trend towards the »soft money« system (Guston 2003). Although it is the best available mechanism (Kostoff 1997), it is not perfect. Peers are not prophets, but ordinary human beings with their own opinions, strengths, and weaknesses (Ehse 2004). Every scientific institution that uses peer review has to deal with the following question: Does the peer review system implemented by my institution fulfil its declared objective to select the best scientific work? As an assessment tool peer review is asked to be *reliable* (is the selection of scientific contributions reliable or is the result purely incidental?), *fair* (are certain groups of applicants favored or at a disadvantage?) and *predictive valid* (do the selection decisions correlate with scientific performance measures subsequent to decision?). We examined in a comprehensive evaluation study the selection process for doctoral and postdoctoral research fellowship recipients followed by the Boehringer Ingelheim Fonds (B.I.F.), a foundation for the promotion of basic research in biomedicine, with regard to these three criteria for professional evaluations (Bornmann 2004 and 2007a; Bornmann and Daniel 2005d).

The data set on which the evaluation is based

We analysed a total of 2,697 applications (1,954 for doctoral and 743 for postdoctoral fellowships) of the years 1985 to 2000. The selection process of the

foundation is configured as follows: Junior scientists submit their fellowship applications to the administrative office (secretariat) of the foundation. The office forwards the application to an independent external reviewer. The external reviewer assesses the applicant, the proposed research project and the institution at which the project will be conducted and in a final statement recommends approval or rejection. In addition to the assessment by an external reviewer, a member of the foundation's staff interviews the applicant personally. Finally, the application, together with the external review and the staff report on the personal interview, is submitted to the B.I.F. Board of Trustees. Seven internationally renowned scientists make up the Board. At each of the three annual Board meetings, the scientists decide on applications.

Criteria used by the board of trustees for the selection of the fellows

The foundation reports that fellowships are awarded to applicants according to the following main criteria: (1) scientific quality as demonstrated by the applicant's achievements to date, (2) the originality of the proposed research project, and (3) the scientific standing of the laboratory where the research will be conducted (according to Hermann Fröhlich, managing director of the B.I.F., see Fröhlich 2001). Using the Boolean probit statistical technique, we examined the multiple conjunctural causation that a fellowship has been awarded only if all three of these criteria were assessed positively by the B.I.F. peer review committee. In agreement with the prescriptive principles of the foundation the results suggest that the B.I.F. approves applications only if all of the three criteria are rated positively (see the results in Bornmann and Daniel 2005b).

Reliability of peer review

Human decisions are classified as reliable when different persons come to the same or similar conclusions. In analysing the reliability of the fellowship selection process at B.I.F., we determined the degree of agreement among the decision-makers. At each of the three annual Board meetings, the seven members of the Board of Trustees decide on applications in three rounds. In the first round of decision-making, some fellowship applications are approved

(rated 'A'), some are rejected (rated 'A-B' and below), and some are earmarked for consideration in the next round (rated 'A-'). In the second and, if necessary, third round, the number of applications approved or dismissed depends on how much funding is still available for the session (Fröhlich 2001). The foundation's secretariat states that the level of controversy in the Trustees' discussion of whether to approve or reject an application increases with the number of rounds. Thus, the round in which the application is approved or rejected should reflect the extent of disagreement among the Trustees: in later rounds, agreement tends to decrease and disagreement increases. The results show that for 76 per cent of the applications, the decisions of the trustees are characterized by agreement, since the decisions on these are reached in the first round. Similar percentages of agreement are reported by other funding institutions (Cicchetti 1991). Decisions are made on 24 per cent of the B.I.F. applications under circumstances in which disagreement more or less prevails (the decisions are reached in the second or third round) (see the results in Bornmann and Daniel 2005d).

In a recently published article, Hargens and Herting (2006) apply the row-column (RC) association model to peer review to analyze the association between two referees' recommendations and an editor's decision at two scholarly journals. The row-column (RC) association model (Goodman 1984) tests whether one (or more) latent dimension can account for the association between recommendations (crossed) and decisions. Using the data on applications for a doctoral or postdoctoral fellowship that were assessed by the B.I.F. by means of the three-stage evaluation process, we tested the extent of the association between reviewers' recommendations (internal and external evaluation) and final decisions on fellowship applications. We show that a single latent dimension is sufficient to account for the association between (internal and external) reviewers' recommendations and the fellowship award decision by the Board. Favorable ratings by the reviewers corresponded with favorable decisions by the Board (and vice versa). This result indicates that the latent dimension underlying reviewers' recommendations and the Board's decisions reflects the merit of an application being evaluated (see the results in Bornmann, Mutz, and Daniel 2007).

Fairness of peer review

Within the first step of our fairness analyses, we investigated some of the most frequently discussed potential sources of bias: applicant's gender, nationality,

discipline and institutional affiliation, i.e. the institution in which the research project is to be carried out. To identify the effect of every single potential source of bias, which could influence decisions of the Board of Trustees, we used multiple logistic regression models. As the foundation had information on the applicants' scientific achievements up to the date of the application, we could include not only the potential sources of bias as independent variables into the statistical analyses, but also the scientific performance of the applicants.

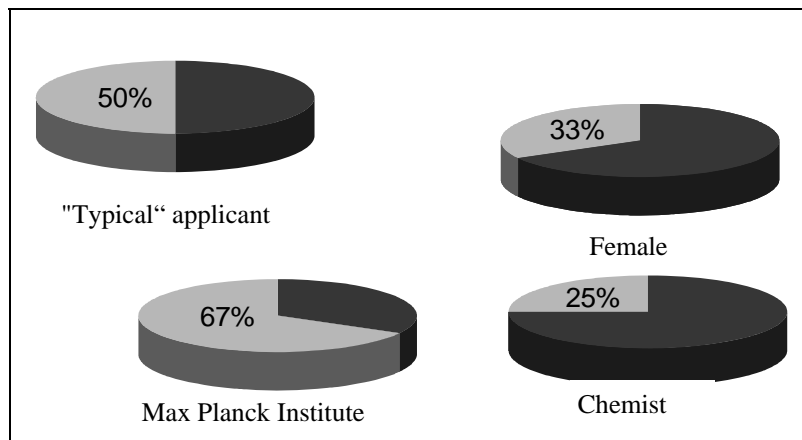
Logistic regression analysis for the applications for a postdoctoral research fellowship showed that none of the examined potential sources of bias has a statistically significant influence on the decisions of the Board of Trustees. With regard to applications for a doctoral fellowship, the applicant's nationality did not statistically significantly affect the Board's decision. However, we detected a statistically significant influence of three variables hypothesized as potential biases: applicant's gender, discipline and intended institutional affiliation. The results on the selection process of the foundation are therefore inconsistent: we found evidence for a gender, discipline and institutional bias in judging applications for doctoral, but not for postdoctoral fellowships. No bias with respect to nationality was found in either group (see the results in Bornmann and Daniel 2005d).

To determine extent and direction of the influence of gender, discipline and intended institutional affiliation on the Board's decisions on doctoral fellowship allocations, we calculated the so-called predicted probabilities of approval and rejection respectively. For the probability calculation, we first simulated a »typical« applicant, based on the average or most common features of all applicants for a doctoral fellowship. This applicant's chance of receiving a scholarship is 50 per cent, as determined by the probability computation (see Figure 1).

If the »typical« applicant is female, the predicted probability of receiving a scholarship decreased from 50 per cent to 33 per cent. Further analyses with classification trees (method: CHAID) pointed out that this lower probability for females of receiving a fellowship is just given among the best qualified applicants (see the results in Bornmann 2006). Moreover, *Figure 1* shows that the impact of the applicant's discipline is still more important than his or her gender: if the applicant is not a biologist, but a chemist, the probability of approval declined from 50 per cent to 25 per cent. The opposite effect is observed for the institution in which the research project will be carried out: with regard to the decision of the Board of Trustees, it is obviously of advantage to choose an institute of the Max Planck Society (Germany) rather than of a German university. This choice increases the probability for approval by 17 percentage points (see the results in Bornmann and Daniel 2004; and our study

of the effects of university prestige and field of study on approval and rejection of fellowship applications, Bornmann and Daniel 2006a). In addition, we investigated the extent to which the foundation's Board of Trustees' practice of reviewing the applications in alphabetic order when making final selection decisions has an influence on the decisions that they make. A statistically significant influence of the postulated bias variable could be observed, but the effect size was small (see the results in Bornmann and Daniel 2005a).

Figure 1. Predicted probabilities for approval (light segments of the circles) and rejection (dark segments of the circles) of an application, taking applicant's gender, discipline or intended institutional affiliation into account (in per cent)



Within the third step of our fairness analyses, we investigated the influence of three attributes of the foundation's external reviewers on their ratings in the selection process: (1) number of applications assessed in the past for the B.I.F. (reviewers' evaluation experience), (2) the reviewers' country of residence and (3) the reviewers' gender. To analyze the reviewers' ratings («award», «maybe award», «no award») in an ordinal regression model the following were considered in addition to the three attributes: (1) the scientific achievements of the fellowship applicants, (2) interaction effects between reviewers' and applicants' attributes, and (3) judgmental tendencies of reviewers. The results of the regression models show no significant effect of the reviewers' attributes on the evaluation of B.I.F. fellowship applications. The ratings of the external reviewers are mainly determined by the applicants' scientific achievement prior to application.

Predictive validity of peer review

In the last part of our evaluation study, we examined the predictive validity of the selection process of the foundation. Assessing the predictive validity of decisions requires a generally accepted criterion for scientific merit. A conventional approach is to use citation counts as a proxy for research impact, since they measure the international impact of the work by individuals or groups of scientists on others (Bornmann and Daniel 2008).

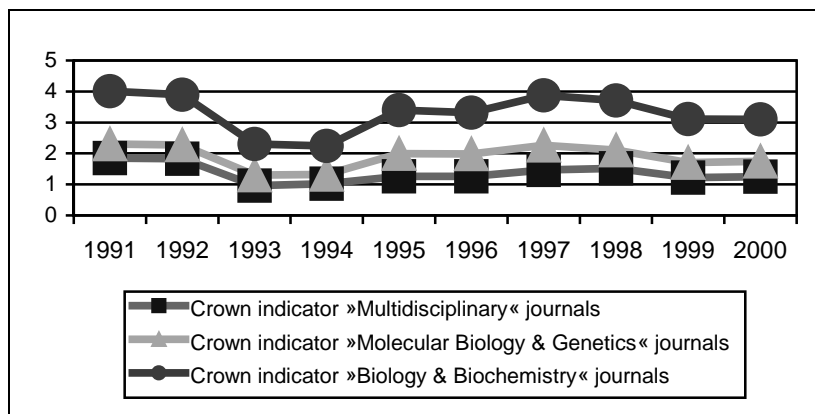
Within the first step of our validity analyses, we examined the predictive validity of the decisions on *doctoral* fellowship applicants. In June 2001, the foundation asked all applicants who had been awarded a fellowship between 1985 and 1995 to send an up-to-date publication list. This list should comprise all publications since the date of approval of the fellowship to December 2000. All in all, 2,039 articles from 120 former fellowship holders could be included in our analyses. By the end of 2001, the 2,039 articles published between 1988 and 2000 had been cited altogether 82,099 times.

Using this bibliometric data, we tried to answer the question whether the judgments of the Board of Trustees correspond with the citation counts for the publications by the Boehringer Ingelheim Fonds fellows. As we described above, at each of the three Board meetings per year, the seven members of the Board decide on applications in three rounds. The round in which the application is approved should reflect the degree of uncertainty among the Trustees. As it can be assumed that greater uncertainty when deciding to award a fellowship is due to lower scientific merit of an application, the round in which a decision was made was used as an indicator for the scientific merit of the application in the analyses of predictive validity described in the following. In later rounds the scientific merits of the applications tends to decrease.

For the analysis of the predictive validity of the B.I.F. peer review process for doctoral fellowship applicants using citation counts as a dependent and the decision round (approved in round 1, 2 or 3) as independent variables, a negative binomial regression model was calculated. The results of the regression model shows that with each later round, the number of expected citations for journal articles by the doctoral fellows decreases by 17 percentage points. This means that the greater the uncertainty of the Trustees when approving a fellowship, the fewer citations can be expected for articles written by the fellow after approval of his or her fellowship. The coefficient for the decision round is statistically significant (see the results in Bornmann and Daniel 2005a), thus the decisions on the applications in three rounds are basically statistically significant.

However, how frequently the publications of a group of scientists (here, doctoral fellowship applicants) have been cited says little on its own (Bornmann and Daniel forthcoming). The assessment of research performance is relative to a frame of reference in which citation counts are interpreted. The comparison with reference standards makes it possible to assign meaning to citation counts and to place them in an adequate context (Bornmann et al. forthcoming). Anthony F. J. van Raan of the Center for Science and Technology Studies (CWTS) in Leiden, Netherlands, recommends a worldwide reference indicator for the bibliometric evaluation of research groups: »Our most important bibliometric indicator, the 'crown indicator', is a trend analysis over a period of, say, eight years, of the number of citations to the entire oeuvre of a research group or institute, normalized to an international field-specific reference value. In this way, we are able to demonstrate whether this group or institute is performing below or above, or even far above the international level of the research field(s) concerned« (van Raan 1999, p. 420).

Figure 2. Crown indicators for articles published by the B.I.F. fellows between 1991 and 2000



Note. If a crown indicator is larger than one, the fellows' publications are more often cited than the 'average' publication in a journal set (e.g. »Molecular Biology and Genetics«)

The calculation of the crown indicators for the B.I.F. applicants for a doctoral fellowship showed that the selection process is highly valid: journal articles published by the fellows are cited considerably more often than the »average« publication in the journal sets »Multidisciplinary«, »Molecular Biology and Genetics«, and »Biology and Biochemistry« (provided by Thomson Scientific,

Philadelphia, PA, USA) (see Figure 2). These sets include journals covering the research fields in which most of the fellows publish (see the results in Bornmann and Daniel 2004, 2005d).

Within the second step of our validity analyses, we examined the decisions of the Board of Trustees on *postdoctoral* fellowship applicants. Here, a citation analysis for articles published previous to the applicants' approval or rejection for a B.I.F. fellowship was conducted. All in all, 1,586 articles (full length articles, letters, notes, communications and reviews) had been published by 397 applicants (64 approved and 333 rejected applicants) previous to their applications to the B.I.F. (on average four articles). Based on a negative binomial regression model, journal articles that had been published by applicants approved for a fellowship award prior to applying for the B.I.F. fellowship award can be expected to have 37 per cent (straight counts of citations) and 49 per cent (complete counts of citations) more citations than articles that had been published by rejected applicants. Furthermore, comparison with international scientific reference values revealed (a) that articles published by successful and non-successful applicants are cited considerably more often than the »average« publication and (b) that excellent research performance can be expected more of successful than non-successful applicants (see the results in Bornmann and Daniel 2006b). The results of both analyses with the data for the postdoctoral fellowship applicants confirm the predictive validity of the B.I.F. selection process.

As the *b* index is particularly well-suited for evaluation of scientists' scientific output, we tested the validity of the B.I.F. funding decisions on applicants for a postdoctoral fellowship also using the *b* index. Hirsch (2005) has proposed the *b* index as a single-number criterion to evaluate the scientific output of a researcher. Hirsch's (2005) index depends on both the number of a scientist's publications and the impact of the papers on peers: »A scientist has index *b* if *b* of his or her N_p papers have at least *b* citations each and the other ($N_p - b$) papers have $\leq b$ citations each« (p. 16569). Even though the *b* index values of approved B.I.F. applicants on average (arithmetic mean and median) are higher than those of rejected applicants (and with this, fundamentally confirm the validity of the funding decisions) (see the results in Bornmann and Daniel 2005c), the distributions of the *b* indices in part overlap with that we categorized as type I error (falsely drawn approval) or type II error (falsely drawn rejection).

In type I error, the B.I.F. Board of Trustees concluded that an applicant had the scientific potential for promotion (and was approved), when he or she actually did not (as reflected in an applicant's low *b* index). In type II error, the Board concluded that an applicant did *not* have the scientific potential for promotion (and was rejected), when he or she actually did (as reflected in a

high *b* index). Based on these definitions, we determined the extent of type I and type II errors in the B.I.F. committee peer review. Approximately one-third of the decisions to award a fellowship to an applicant show a type I error, and about one-third of the decisions not to award a fellowship to an applicant show a type II error (see the results in Bornmann and Daniel 2007).

Finally, we would like to mention that although the results of the evaluation of the B.I.F. funding decisions point to type I and type II errors in the Board's decisions in a part of the applications, approved applicants, as compared to rejected applicants, had on average published more papers prior to applying for the fellowship, and their later publications had greater impact. While it would certainly be desirable to completely eliminate both error types in the B.I.F. peer review procedure, it simply cannot be done. In fact, reducing one cause for one error type (e.g., by increasing the approval rate) automatically increases the risk for the other error type.

Additional studies could further substantiate the validity of the Foundation's selection process by analyzing other success criteria, such as the applicants' professional career (Teichler 1991) or by statistics on third-party funds and patents (Hornbostel 1991) of former fellowship holders. This would also provide information on the interrelation between different indicators of success.

Discussion

In our comprehensive evaluation study, we investigated the decisions for awarding long-term fellowships to doctoral and postdoctoral researchers as practiced by the B.I.F. The secretariat of the foundation presented each of the Trustees with our core findings with the dedication: »Nothing is so good that it can't be made even better« (Fröhlich 2004). In fact, our results on the reliability and predictive validity of the foundation's peer review process indicate that the process is generally a credible method for *ex ante* evaluation of fellowship applications (Smaglik 2004). But our findings on the fairness of the process show that there are also problems with peer review (especially, the applicant's gender as a potential source of bias).

The results of our study were thoroughly discussed by the B.I.F. Board of Trustees and the foundation continued to monitor its selection process closely. This allowed the B.I.F. to see a considerable increase in female applicants and scholars in the next few years, with nearly 50 per cent of the 2006 PhD scholarships awarded to women. But according to Hermann Fröhlich, managing

director of the B.I.F., the growing number of young women participating and succeeding in one of the most competitive selection processes for fellowships may be due to social change. And as the B.I.F. evaluates young researchers and their projects at the earliest possible phase of the scientific career, its figures may indicate that larger numbers of women have started to reach for the top in science (Bornmann 2007b).

However, despite its flaws, having scientists judge each other's work is widely considered to be the »least bad way« to weed out weak research proposals and improve promising ones. Therefore *ex ante* peer review should be used for the evaluation of fellowship 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 (Daniel 2005; Daniel, Mittag, and Bornmann 2007).

References

- Bornmann, L. (2004): *Stiftungspropheten in der Wissenschaft. Zuverlässigkeit, Fairness und Erfolg des Peer-Review*. Münster: Waxmann.
- Bornmann, L. (2006): »Peer-Review zur Auswahl von Forschungsstipendiaten. Eine Analyse der Fairness und prognostischen Validität des Auswahlprozesses mittels CHAID und GLM«. In: *Empirische Pädagogik*, Vol. 20, No. 4, pp. 347-368.
- Bornmann, L. (2007a): »Bestenauswahl mit Peer Review – Die Vergabe von Forschungsstipendien des Boehringer Ingelheim Fonds«. In: *Forschung und Lehre*, Vol. 14, no. 6, pp. 332-333.
- Bornmann, L. (2007b): »Bias Cut. Women, it Seems, Often Get a Raw Deal in Science - so How can Discrimination be Tackled?«. In: *Nature*, 445(7127), p. 566.
- Bornmann, L., and Daniel, H.-D. (2004): *Reliability, Fairness and Predictive Validity of Committee Peer Review. Evaluation of the Selection of Post-Graduate Fellowship Holders by the Boehringer Ingelheim Fonds*. In: *B.I.F. Futura*, 19, pp. 7-19
- Bornmann, L., and Daniel, H.-D. (2005a): »Committee Peer Review at an International Research Foundation: Predictive Validity and Fairness of Selection Decisions on Post-Graduate Fellowship Applications«. In: *Research Evaluation*, Vol. 14, no. 1, pp. 15-20.
- Bornmann, L., and Daniel, H.-D. (2005b): »Criteria Used by a Peer Review Committee for Selection of Research fellows – a Boolean Probit Analysis«. In: *International Journal of Selection and Assessment*, Vol. 13, no. 4, pp. 296-303.
- Bornmann, L., and Daniel, H.-D. (2005c): »Does the h-index for Ranking of Scientists Really Work?«. In: *Scientometrics*, Vol. 65, no. 3, pp. 391-392.
- Bornmann, L., and Daniel, H.-D. (2005d): »Selection of Research Fellowship Recipients by Committee Peer Review. Analysis of Reliability, Fairness and Predictive Validity of Board of Trustees' decisions«. In: *Scientometrics*, Vol. 63, no. 2, pp. 297-320.

- Bornmann, L., and Daniel, H.-D. (2006a): »Potential Sources of Bias in Research Fellowship Assessments. Effects of University Prestige and Field of Study on Approval and Rejection of Fellowship Applications«. In: *Research Evaluation*, Vol. 15, No. 3, pp. 209-219.
- Bornmann, L., and Daniel, H.-D. (2006b): »Selecting Scientific Excellence Through Committee Peer Review – a Citation Analysis of Publications Previously Published to Approval or Rejection of Post-doctoral Research Fellowship Applicants«. In: *Scientometrics*, Vol. 68, No. 3, pp. 427-440.
- Bornmann, L., and Daniel, H.-D. (2007): »Convergent validation of peer review decisions using the h index: extent of and reasons for type I and type II errors«. In: *Journal of Informetrics*, Vol. 1, No. 3, pp. 204-213.
- Bornmann, L., and Daniel, H.-D. (2008): »What do Citation Counts Measure? A Review of Studies on Citing Behavior«. In: *Journal of Documentation*, Vol. 64, No. 1, pp. 45-80.
- Bornmann, L., Mutz, R., and Daniel, H.-D. (2007): »Row-column (RC) association model applied to grant peer review«. In: *Scientometrics*, Vol. 73, No. 2, pp. 139-147.
- Bornmann, L., Mutz, R., Neuhaus, C., and Daniel, H.-D. (in press): Use of Citation Counts for Research Evaluation: Standards of Good Practice for Analyzing Bibliometric Data and Presenting and Interpreting Results. In: *Ethics in Science and Environmental Politics*.
- Cicchetti, D. V. (1991): »The Reliability of Peer Review for Manuscript and Grant Submissions: a Cross-disciplinary Investigation«. In: *Behavioral and Brain Sciences*, Vol. 14, No. 1, pp. 119-135.
- Daniel, H.-D. (2005): »Publications as a Measure of Scientific Advancement and of Scientists' Productivity«. In: *Learned Publishing*, 18, pp. 143-148.
- Daniel, H.-D., Mittag, S., and Bornmann, L. (2007): »The Potential and Problems of Peer Evaluation in Higher Education and Research«. In: Cavalli, A. (ed.): *Quality Assessment for Higher Education in Europe*. London, UK: Portland Press, pp. 71-82.
- Ehse, I. (2004): »By Scientists, for Scientists. The Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – and how it Functions«. In: *B.I.F. Futura*, 19, pp. 170-177.
- Fröhlich, H. (2001): »It All Depends on the Individuals. Research Promotion – a Balanced System of Control«. In: *B.I.F. Futura*, 16, pp. 69-77.
- Fröhlich, H. (2004): »In the hands of social researchers«. In: *B.I.F. Futura*, 19, pp. 19-23.
- Goodman, L. A. (1984): *The Analysis of Cross-classified Data Having Ordered Categories*. Cambridge, MA, USA: Harvard University Press.
- Guston, D. H. (2003): »The Expanding Role of Peer Review Processes in the United States«. In: Shapira, P. and Kuhlmann, S. (eds.): *Learning from Science and Technology Policy Evaluation. Experiences from the United States and Europe*. Cheltenham, UK: Edward Elgar, pp. 81-97.
- Hargens, L. L., and Herting, J. R. (2006): »Analyzing the Association Between Referees' Recommendations and Editors' Decisions«. In: *Scientometrics*, Vol. 67, No. 1, pp. 15-26.
- Hirsch, J. E. (2005): »An Index to Quantify an Individual's Scientific Research Output«. In: *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 102, No. 46, pp. 16569-16572.

- Hornbostel, S. (1991): »Drittmittel im Fach Physik – ein Indikator für Forschungsleistungen?« In: *Physikalische Blätter*, 2, pp. 123-125.
- Kostoff, R. N. (1997): »The Principles and Practices of Peer Review«. In: *Science and Engineering Ethics*, Vol. 3, No. 1, pp. 19-34.
- Smaglik, P. (2004): »Up for review«. In: *Nature*, Vol. 430, 591.
- Teichler, U. (1991): »Evaluation of the EC Training Fellowship Program Based on a Fellows Questionnaire Survey«. In: *Scientometrics*, Vol. 21, No. 3, pp. 343-365.
- van Raan, A. F. J. (1999): »Advanced Bibliometric Methods for the Evaluation of Universities«. In: *Scientometrics*, Vol. 45, No. 3, pp. 417-423.