Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions

LUTZ BORNMANN,[a] HANS-DIETER DANIEL[b,c]

[a] *Swiss Federal Institute of Technology Zurich (ETH Zurich), Professorship for Social Psychology and Research on Higher Education, Zurich(Switzerland)*
[b] *University of Zurich, Evaluation Office, Zurich (Switzerland)*
[c] *Professor for Social Psychology and Research on Higher Education, ETH Zurich (Switzerland)*

In science, peer review is the best-established method of assessing manuscripts for publication and applications for research fellowships and grants. However, the fairness of peer review, its reliability and whether it achieves its aim to select the best science and scientists has often been questioned. The paper presents the first comprehensive study on committee peer review for the selection of doctoral (Ph.D.) and post-doctoral research fellowship recipients. We analysed the selection procedure followed by the Boehringer Ingelheim Fonds (B.I.F.), a foundation for the promotion of basic research in biomedicine, with regard to the reliability, fairness and predictive validity of the procedure – the three quality criteria for professional evaluations. We analysed a total of 2,697 applications, 1,954 for doctoral and 743 for post-doctoral fellowships. In 76% of the cases, the fellowship award decision was characterized by agreement between reviewers. Similar figures for reliability have been reported for the grant selection procedures of other major funding agencies. With regard to fairness, we analysed whether potential sources of bias, i.e., gender, nationality, major field of study and institutional affiliation, could have influenced decisions made by the B.I.F. Board of Trustees. For post-doctoral fellowship applications, no statistically significant influence of any of these variables could be observed. For doctoral fellowship applications, we found evidence of an institutional, major field of study and gender bias, but not of a nationality bias. The most important aspect of our study was to investigate the predictive validity of the procedure, i.e., whether the foundation achieves its aim to select as fellowship recipients the best junior scientists. Our bibliometric analysis showed that this is indeed the case and that the selection procedure is thus highly valid: research articles by B.I.F. fellows are cited considerably more often than the "average" paper (average citation rate) published in the journal sets corresponding to the fields "Multidisciplinary", "Molecular Biology & Genetics", and "Biology & Biochemistry" in Essential Science Indicators (ESI) from the Institute for Scientific Information (ISI, Philadelphia, Pennsylvania, USA). Most of the fellows publish within these fields.

Introduction

Peer Review is central to the process of modern science (KOSTOFF, 1997, p. 32). It is the most important method for the assessment of grant applications, manuscripts submitted for publication in journals and applications for research fellowships (JAYASINGHE et al., 2003). As "gatekeepers" of science, the task of peers or colleagues asked to evaluate applications or manuscripts is to recommend only those that meet the highest of scientific standards. In the selection of research projects, "the peer review system is asked to be highly sensitive and highly selective at the same time. A sensitive review system would detect the merit in every worthwhile proposal, while a selective system would filter out all projects of dubious quality or significance. In effect, a sensitive system captures the signal, however faint, while a selective one removes the noise, however innocuous" (HACKETT & CHUBIN, 2003, pp. 16-17). Alternatives to peer review – such as earmarking funds for specific institutions, using formula to allocate resources and using a lottery for awarding funds – violate the principle that applications should be evaluated and recognized on the basis of scientific merit (EISENHART, 2002; HACKETT & CHUBIN, 2003; INCE, 1991; RENNIE, 2003).

POLANYI (1966) regards peer review as the embodiment of the principle of mutual control, which fosters judgements with respect to the novelty, accuracy and relevance of research results. Proponents of the system hold that peer review is more effective than any other known instrument for self-regulation in promoting the critical selection that is so crucial to the evolution of scientific knowledge. "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" (EISENHART, 2002, p. 241). Thus, the producers of science, the specialists, become the gatekeepers of science (MCCLELLAN, 2003, p. 95).

Although there is evidence that peer review improves the quality of the reporting of research results (GOODMAN et al., 1994; PIERIE et al., 1996), critics of peer review argue that (i) reviewers rarely agree on whether or not to recommend that a manuscript be published or a research fellowship be awarded, thus making for poor reliability of the peer review procedure; (ii) reviewers' recommendations 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 procedure lacks predictive validity, since there is little or no relationship between the reviewers' judgements and the subsequent usefulness of the work to the scientific community, as indicated by the frequency of citations of the work in later scientific papers (for criticism on peer review see ABATE, 1995; FINN, 2002; HORROBIN, 2001; LANGFELDT, 2004; MORAN, 1998; ROSS, 1980; ROY, 1985).

The empirical research on peer review has dealt mainly with the assessment of journal submissions (CAMPANARIO, 1998a; CAMPANARIO, 1998b; OVERBEKE & WAGER, 2003; WELLER, 2002) and grant applications (BORNMANN & DANIEL, 2003;

Demicheli & Pietrantonj, 2004; Wessely, 1998; Wood & Wessely, 2003). The selection by committee peer review of post-graduate researchers (doctoral (Ph.D.) and post-doctoral) for scholarship and fellowship grants has received little attention. A few years ago, the Boehringer Ingelheim Fonds (for more information, see the B.I.F. Web site at http://www.bifonds.de), a foundation for the promotion of basic research in biomedicine located in Heidesheim, Germany, agreed to have us conduct an independent external evaluation of its selection procedure for awarding doctoral and post-doctoral fellowships (Bornmann, 2004; Bornmann & Daniel, 2004). Our evaluation study aimed to answer two questions: (i) does the peer review system fulfil its declared objective to select the best junior scientists for fellowships?; (ii) are the main criticisms raised against peer review as outlined above justified? Here we present the most important results of our investigation, which is the most comprehensive study on selection of post-graduate fellowship recipients conducted to date.

The data set on which the evaluation is based

The archive of the administrative office of the Boehringer Ingelheim Fonds keeps the files of the majority of the fellowship applications. The files contain curriculum vitae, reviews, references, appraisals, protocols of the decision-making Board meetings and other documents. All in all, 2,697 applications received by the foundation between 1985 and 2000 were available for analysis: 1,954 applications for doctoral fellowships (72%) and 743 applications for post-doctoral fellowships (28%). The number of applications for the latter is much lower, because the foundation discontinued post-doctoral fellowships in 1995.

The selection procedure of the Boehringer Ingelheim Fonds

Junior scientists submit their fellowship applications to the administrative office (secretariat) of the foundation, which checks that the applicant and proposed project fulfil the formal requirements and that all required documents have been submitted.[*] Once the formal criteria have been met, the office forwards each application to an independent external reviewer. On the basis of predetermined criteria, the reviewer assesses the applicant, the proposed research project and the institution in which the project is to be conducted and recommends approval or rejection.

---

[*] Dr. Hermann Fröhlich, managing director of the Boehringer Ingelheim Fonds since 1990, provides a detailed description of the selection procedure (Fröhlich, 2001).

Table 1 shows the ratings given by the external reviewers for applications received from 1985 to 2000.[*] The reviewers recommended awarding foundation fellowships for 62% of the applications for a doctoral fellowship and 59% of the applications for a post-doctoral fellowship. In both groups, the external reviewers recommended "no award" for about 20% of the applications.

Table 1. Ratings given by the external reviewers to applications for doctoral and post-doctoral fellowships (in percent)

Rating	Applications for doctoral research fellowships ($n = 1,490$)	Applications for post-doctoral research fellowships ($n = 491$)
Award	62	59
Possible award	17	19
No award	21	22
Total	100	100

In addition to the assessment by an external reviewer, a member of the foundation's staff also examines the application, interviews the applicant personally and submits a detailed report. The staff member rates the application as follows: (i) "definite award", (ii) "award", (iii) "possible award" or (iv) "no award".[**]

Table 2. Ratings given by the staff of the foundation to applications for doctoral and post-doctoral fellowships (in percent)

Rating	Applications for doctoral research fellowships ($n = 1,920$)	Applications for post-doctoral research fellowships ($n = 704$)
Definite award	10	8
Award	33	27
Possible award	28	27
No award	29	38
Total	100	100

Table 2 shows the ratings of all applications for doctoral and post-doctoral fellowships by foundation staff between 1985 and 2000. In both groups, about 10% of

---

[*] Since the reviewers themselves did not use a rating scale, two experts of the Centre for Research on Higher Education and Work (Kassel, Germany) independently rated all reviews afterwards according to the scale shown in Table 1. The reliability of the two experts' ratings is very high (weighted kappa coefficient = 0.96).
[**] The interviewers use a rating scale.

the applications were strongly recommended for a "definite award" and about 30% were recommended for an "award". Twenty-nine percent of the applications for a doctoral and 38% of the applications for a post-doctoral research fellowship were recommended for "no award".

Finally, the applications, together with the external reviews and the staff ratings along with reports on the personal interview, are submitted to the Board of Trustees. Seven internationally renowned scientists make up the Board. The Board convenes three times a year to make approval or rejection decisions after discussing each individual application in detail. From 1985 to 2000, the Board approved 25% of the applications for doctoral fellowships and about 20% of the applications for post-doctoral research fellowships. A comparison of these percentages with the external reviewers' recommendations (Table 1) and the foundation's staff recommendations (Table 2) reveals that both the external reviewers and foundation staff more frequently recommended approval than the Board of Trustees did. About 65% of applications rated "award" by the reviewers and about 50% of applications rated "definite award" or "award" by the foundation's staff did not receive a research fellowship in the end.*

In a study on panel peer review at the National Science Foundation (Arlington, Virginia, USA), KLAHR (1985) found similar results: ratings of the ad hoc reviewers (the external reviewers) "are more 'lenient' than the panel ratings" (p. 151). KLAHR (1985) considers the following causes for the discrepancies: "The ad hoc [the external] reviewers may have more technical proficiency, a better sense of what can realistically be accomplished in the area, and greater familiarity with the track record of the principal investigator. However, the ad hoc reviewers are at a disadvantage when it comes to making a quantitative rating of the proposal. First of all, they are generally unfamiliar with the ratings that get translated into decisions. Second, they do not have the same sense of scarce resources that the panellists do" (p. 152).

Reliability, fairness and predictive validity of the Boehringer Ingelheim Fonds' selection procedure

The Board of Trustees of the Boehringer Ingelheim Fonds has the difficult task of assessing the scientific merit of the applicants and their research proposals and selecting the best junior scientists for fellowships. We investigated the extent to which the Board was able to accomplish this challenging objective between 1985 and 2000. The committee peer review procedure of the foundation was examined with regard to the

---

* An overview from the UNITED STATES GENERAL ACCOUNTING OFFICE (1999, Washington, DC, USA) of peer review practices in twelve federal science agencies found that all of the agencies "use a combination of external and internal reviewers with subject matter expertise" (p. 6).

quality criteria for professional evaluations: reliability (is the selection of fellowship recipients reliable or is the result purely incidental?), fairness (are certain groups of applicants favoured or at a disadvantage?) and predictive validity (does the procedure fulfil the objective to select the best junior scientists?).

Reliability of committee 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 procedure 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, p. 76). 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. Table 3 shows that for 76% of the applications, the decisions of the trustees are characterized by agreement, since the decisions on these are reached in the first round. Decisions are made on 24% of the applications under circumstances in which disagreement more or less prevails.

Table 3. Number of decisions made by the Board of Trustees in each of three rounds
(in percent; $n = 2,524$)

First round	Second round	Third round
76% ($n = 1,905$)	16% ($n = 394$)	8% ($n = 225$)

Agreement ⟶ Disagreement

Fairness of committee peer review

Journal submissions or fellowship grant applications are supposed to be judged solely on the basis of their scientific merit. Personal characteristics and specific attributes of authors or applicants, such as applicants' gender or nationality, should not influence the procedure; otherwise the fairness of the procedure is at risk. In a review of

the literature, ROSS (1980) and SHARP (1990) refer to 16 potential sources of bias,[*] OWEN (1982) reports 25 potential sources of bias and HOJAT et al. (2003), PRUTHI et al. (1997) and WOOD & WESSELY (2003) list about ten. In an overview of the state of research on peer review, BORNMANN & DANIEL (2003, pp. 211-216) review the research on three potential sources of bias.

In the framework of our present study, we investigated some of the most frequently examined potential sources of bias: the applicant's gender, nationality (German or foreign), major field of study (biology, biochemistry, chemistry or medicine) and institutional affiliation, meaning the institution in which the research project is to be carried out: German university, European Molecular Biology Laboratory (EMBL, Heidelberg, Germany), Helmholtz Association of National Research Centres (HGF, Bonn, Germany), Max Planck Society (MPS, Munich, Germany) or Wissenschafts-gemeinschaft Gottfried Wilhelm Leibniz (WGL, Bonn, Germany) (see Table 4, bottom).

To identify the effect of every single potential source of bias that could influence the Board of Trustees' decisions, we used multiple logistic regression models (HOSMER & LEMESHOW, 2000). These models are appropriate for the analysis of dichotomous (or binary) responses. Dichotomous responses arise when the outcome is presence or absence of an event (RABE-HESKETH & EVERITT, 2004, p. 98). In the case of the Boehringer Ingelheim Fonds, the binary response is coded 1 for approval and 0 for rejection of an application. As the foundation had information on the applicants' scientific achievements up to the date of their fellowship applications, we could therefore include not only the potential sources of bias, but also the scientific performance of the applicants as independent variables in the statistical analyses. We were thus able to distinguish between the influence of the applicants' achievements to date and the potential sources of bias on the decisions of the Board. COLE & FIORENTINE (1991) call this proceeding in statistical bias analysis the "control variable approach" (p. 216).[**]

---

[*] Bias is defined as the influence of variables reflecting something other than the applicant's scientific merit. Such variables could be the applicant's age, gender, institutional affiliation or research field (GODLEE & DICKERSIN, 2003, p. 92; ROSS, 1980, pp. 78-79).

[**] The logistic regression model assumes a linear relationship between the natural logarithm of the probability of success (here the granting of a fellowship) and the interval independent variables (HOSMER & LEMESHOW, 2000). Before interpreting regression models, therefore, it is necessary to test the validity of this assumption, which can be done using, for example, CLEVELAND's (1979) locally weighted scatterplot smoother. The tests of the independent variables in Table 4 showed that four variables violate the assumption of linearity. Following recommendations by MOSTELLER & TUKEY (1977), the variables – depending on the type of violation – were entered into the model estimates as logarithmic or quadratic transformations.

Table 4. Description of the independent variables

Independent variable	Applicants for doctoral fellowships ($n = 1,022$)		Applicants for post-doctoral fellowships ($n = 134$)	
	Values	Mean value or percent of value '1'	Values	Mean value or percent of value '1'
Year of Board of Trustees' meeting	1985 → 2000	1994.8	1990 → 1995	1992.7
Scientific performance indicators				
Applicant's age at the time of the final degree	22 → 34	25.9	–	–
Applicant's age at the time of receiving Ph.D.	–	–	25 → 38	28.7
Final grade (.88=highest grade)	0.88 → 4.0	1.4	–	–
Final grade (Ph.D., 1=highest grade)	–	–	1 → 3	1.8
Applicant's mobility during education (1=mobile, 0=immobile)	0 → 1	55%	0 → 1	96%
Number of recommendation letters	0 → 3	1.9	1 → 3	2.3
Rating by external reviewers (1=award, 2=possible award, 3=no award)	1 → 3	1.6	1 → 3	1.6
Rating by members of the foundation's staff (1=definite award, 2=award, 3=possible award, 4=no award)	1 → 4	2.7	1 → 4	3.0
Number of journal articles published by applicant by the time of application	–	–	0 → 23	3.7
Potential sources of bias				
Gender (1=female, 0=male)	0 → 1	42%	0 → 1	37%
Nationality (1=foreign, 0=German)	0 → 1	3%	0 → 1	6%
Applicant's major field of study:				
- Biology (=1, 0=other field of study)	0 → 1	60%	0 → 1	43%
- Biochemistry (=1, 0= other field of study)	0 → 1	14%	0 → 1	6%
- Chemistry (=1, 0= other field of study)	0 → 1	10%	0 → 1	13%
- Medicine (=1, 0= other field of study)	0 → 1	2%	0 → 1	33%
- Further fields of study (reference category)	0 → 1	14%	0 → 1	5%
Institution where the research project will be conducted:[*]				
- University (=1, 0=other institution)	0 → 1	52%	0 → 1	15%
- EMBL (=1, 0= other institution)	0 → 1	2%	0 → 1	0%
- HGF (=1, 0= other institution)	0 → 1	4%	0 → 1	0%
- MPS (=1, 0= other institution)	0 → 1	9%	0 → 1	4%
- WGL (=1, 0= other institution)	0 → 1	3%	0 → 1	0%
- Further institutions (reference category)	0 → 1	30%	0 → 1	81%

[*] Due to the variety of German university and non-university research institutions for which the applicants submitted a fellowship application, only those institutions with a large enough sample size were entered into the analyses as independent variables: university, EMBL, HGF, MPS and WGL. The remaining institutions were grouped together as "further institutions", forming a reference category in the logistic regression analyses.

The following scientific performance indicators essentially comprise the criteria for approval and rejection of an application in the selection procedure of the Boehringer Ingelheim Fonds: (i) for applicants for a doctoral fellowship: age at the time of the final degree, final grades, mobility during education (mobile or immobile), the number of recommendation letters and the ratings by the external reviewers ("award", "possible award", "no award") and members of the foundation's staff ("definite award", "award", "possible award", "no award"); (ii) for applicants for a post-doctoral research fellowship: age at the time of receiving Ph.D., grades, mobility during education, the number of letters of recommendation, the number of published journal articles by the time of application as well as the ratings by the external reviewers and the foundation's staff (see Table 4, top).

Table 5 and Table 6 show the results of the multiple regression analyses for predicting the Board of Trustees' decisions on awarding fellowships from scientific performance indicators and potential sources of bias. The results of the likelihood ratio tests are χ^2 (18, $n = 1,022$) = 72.3, $p < 0.001$ (applicants for doctoral fellowships) and χ^2 (16, $n = 134$) = 44.6, $p < 0.001$ (applicants for post-doctoral fellowships).[*] As the p values for the tests are significant at the $\alpha = 0.001$ level, we reject the null hypothesis and conclude that at least one and perhaps all odds ratios in the models are different from zero.

Table 5 shows the results of the model estimates predicting the Board of Trustees' decisions on post-doctoral fellowships.

The results of the analyses for scientific performance indicators show that number of journal articles published and the rating by members of the foundation's staff had a significant effect on the Board's decisions on post-doctoral fellowships: the odds of approval of a post-doctoral fellowship increase with each published journal article by the time of application. To determine the extent and direction of the influence of the ratings by members of the foundation's staff on the Board's decisions, we calculated so-called predicted probabilities of approval (CONROY, 2002). The results show that the predicted success rate of approval of post-doctoral fellowships is 62% for "definite award," is about a third for "award" and falls to about 1% for "no award".

---

[*] The logistic regression models had to be calculated with reduced sample sizes, as only those cases could be included in the statistical analyses that had no missing values for the variables entered into the model. As a result, 52% ($n = 1,022$) of the applicants for doctoral fellowships and 18% ($n = 134$) of the applicants for post-doctoral fellowships could be included. Although it is possible to include cases with missing data in the analysis using imputation methods (MANDER & CLAYTON, 1999; RUBIN & SCHENKER, 1986) such as provided by the statistical package Stata (STATACORP., 2003), the parameter estimates fluctuate depending on the imputation method or – in some imputation methods – according to the number of imputations performed (RUBIN, 1987; SCHAFER, 2000). Because the parameters estimated in this way vary highly and in part can hardly be replicated, no imputation methods were used for the model estimates.

Table 5. Regression analysis predicting Board of Trustees' decisions on post-doctoral fellowships from scientific performance indicators and potential sources of bias ($n = 134$)

Independent variable	Odds ratio	Standard error	p value
Year of Board of Trustees' meeting (squared)	1.00	0.00	0.627
Scientific performance indicators			
Applicant's age at the time of receiving Ph.D.	0.82	0.17	0.334
Final grade (1=highest grade)	0.69	0.40	0.517
Applicant's mobility during education (1=mobile, 0=immobile)	2.17	4.33	0.698
Number of recommendation letters	0.73	0.43	0.589
Rating by external reviewers (1=award, 2=possible award, 3=no award)	0.23	0.19	0.070
Rating by members of the foundation's staff (1=definite award, 2=award, 3=possible award, 4=no award)	0.34	0.11	0.001
Number of journal articles published by the time of application (logarithmic)	1.21	0.11	0.047
Potential sources of bias			
Gender (1=female, 0=male)	0.74	0.53	0.672
Nationality (1=foreign, 0=German)	0.92	1.53	0.961
Applicant's major field of study:			
- Biology (=1, 0=other field of study)	9.73	15.30	0.148
- Biochemistry (=1, 0=other field of study)	6.87	12.37	0.285
- Chemistry (=1, 0=other field of study)	2.05	3.53	0.676
- Medicine (=1, 0=other field of study)	6.06	9.90	0.271
Institution where the research project will be conducted:			
- University (=1, 0=other institution)	2.38	2.22	0.355
- EMBL (=1, 0=other institution)	–*	–*	–*
- HGF (=1, 0=other institution)	–*	–*	–*
- MPS (=1, 0=other institution)	2.98	5.81	0.576
- WGL (=1, 0=other institution)	–*	–*	–*

* Number of cases is too small for statistical analyses.

No statistically significant influence was found for the other scientific performance indicators that were included in the multiple regression analysis (applicant's age at the time of receiving Ph.D., final grade, applicant's mobility during education, number of recommendation letters and rating by the external reviewers), and no statistically

significant influence was found for the potential sources of bias examined (applicant's gender, nationality, major field of study and institution in which the research is to be conducted).

Table 6. Regression analysis predicting Board of Trustees' decisions on doctoral fellowships from scientific performance indicators and potential sources of bias ($n = 1,022$)

Independent variable	Odds ratio	Standard error	p value
Year of Board of Trustees' meeting (squared)	1.00	0.00	0.000
Scientific performance indicators			
Applicant's age at the time of the final degree	0.85	0.06	0.021
Final grade (0.88=highest grade, logarithmic)	0.19	0.09	0.000
Applicant's mobility during education (1=mobile, 0=immobile)	1.40	0.32	0.150
Number of recommendation letters	0.84	0.12	0.206
Rating by external reviewers (1=award, 2=possible award, 3=no award)	0.42	0.08	0.000
Rating by members of the foundation's staff (1=definite award, 2=award, 3=possible award, 4=no award)	0.24	0.03	0.000
Potential sources of bias			
Gender (1=female, 0=male)	0.49	0.10	0.001
Nationality (1=foreign, 0=German)	1.42	0.83	0.546
Applicant's major field of study:			
- Biology (=1, 0=other field of study)	1.17	0.40	0.656
- Biochemistry (=1, 0=other field of study)	1.41	0.56	0.381
- Chemistry (=1, 0=other field of study)	0.39	0.18	0.045
- Medicine (=1, 0=other field of study)	0.72	0.68	0.725
Institution where the research project will be conducted:			
- University (=1, 0=other institution)	0.93	0.25	0.800
- EMBL (=1, 0=other institution)	0.68	0.40	0.510
- HGF (=1, 0=other institution)	0.68	0.35	0.446
- MPS (=1, 0=other institution)	1.91	0.61	0.046
- WGL (=1, 0=other institution)	1.57	0.89	0.430

The results presented in Table 6 for predicting Board of Trustees' decisions on doctoral fellowships show that four of the six scientific performance indicators had a significant effect in the expected direction: the Board was more likely to award a

doctoral fellowship the younger the applicants were at the time of the final degree, the higher their final grades, and the higher the ratings by the external reviewers and the members of the foundation's staff. No statistically significant influence was found only for applicant's mobility during education and number of recommendation letters.

As to potential sources of bias, the applicant's nationality did not have a statistically significant effect on the Board's decision to approve a doctoral fellowship. However, we detected a statistically significant influence of three variables hypothesized as potential biases: applicant's gender, major field of study (chemistry) and research institution affiliation (MPS). The calculation of the predicted probabilities (Conroy, 2002) shows that it is obviously an advantage if the applicant is affiliated with an institute of the Max Planck Society (46%) rather than another research institution (10%). The choice of a Max Planck Institute for conducting the research increases the predicted probability for approval by 36 percentage points. The opposite effect was found for female applicants and applicants working in the field of chemistry: the predicted success rate of approval of doctoral fellowships for applicants working in the field of chemistry (6% predicted success rate) is only approximately half as high as the predicted success rate of approval (12%) for applicants working in other major fields of study. The same was found for the success rate of approval for women (7%) compared to male applicants (16%). All in all, the results of the probability calculations for applicants for doctoral fellowships indicate that the Board of Trustees tends to rate particular applicant groups more highly than others.[*]

To sum up, the results on the foundation's selection procedure are inconsistent. We found evidence for a gender, major field of study and institutional bias in approving applications for doctoral, but not for post-doctoral fellowships. No bias with respect to nationality was found in either group.

---

[*] To demonstrate extent and direction of the influence of gender, discipline and intended institutional affiliation on the Board's decisions on doctoral fellowship allocations in another analysis Bornmann & Daniel (2004, pp. 10-11) simulated a "typical" applicant, based on the average or most common features of all applicants for a doctoral fellowship. The "typical" applicant completed his university degree at the age of 26 with a final grade of 1.4 (best grade is 1.0). He attended more than one university during his education. In addition, he could submit two letters of recommendation with his application. Both the external reviewer and the foundation's staff recommended him for an award. He is male, of German nationality and his first degree is in biology. He will pursue his research project at a German university. This applicant's chances of receiving a scholarship are 50%, as determined by the probability computation. If the "typical" applicant is not male, but female, the predicted probability of receiving a scholarship decreased from 50% to 33%. The impact of the applicant's discipline is still more important: if the applicant is not a biologist, but a chemist, the probability of approval declined from 50% to 25%. 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.

Predictive validity of committee peer review

In the third part of our study, we examined the predictive validity of the foundation's selection procedure, that is, whether indeed the "best" junior scientists are selected to receive fellowships. 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 (COLE, 2000, p. 293).

Only fellowship recipients and not non-selected applicants were included in the assessments of predictive validity, because "criterion data for rejected applicants are difficult to obtain and difficult to interpret, even when available; those accepted are no longer comparable to those rejected because two groups have had different experiences" (CHAPMAN & MCCAULEY, 1994, p. 428). According to Gerhard Sonnert,[*] B.I.F. fellowships clearly have a dual function. They reward previous excellence (i.e., they are given to the "best" applicants, who are selected according to merit criteria), but they also afford the successful applicants resources that might enable them to do excellent scientific work in their future careers. With reference to Robert Merton's concept of "self-fulfilling prophecy" (MERTON, 1948), one could argue that the fellowships of the B.I.F. give the fellows such an advantage in training, opportunities, prestige, self-confidence, and so on that they later become superior scientists because of the fellowship, not because they were particularly promising at the point of application.[**] Rather than picking the best scientists, the selection committee might, in this view, create them (see also COLE & COLE, 1967; HAGSTROM, 1965; MERTON, 1968).

In June 2001, the foundation asked former fellowship recipients who had received their fellowships between 1985 and 1995 to submit an up-to-date publication list of all works published from the date of approval of the fellowship up to December 2000. Of 433 B.I.F. fellows, 225 (52%) responded and sent in their list of publications. The foundation secretariat determined whether each of the 225 fellows was working in an academic institution (publicly funded research), or in industry or as (for example) a medical doctor, patent attorney or journalist. Sixty-three percent (141 fellows) of the 225 former scholarship holders had worked exclusively in academic institutions.

[**] In particular, since the Boehringer Ingelheim Fonds organizes different seminars for its fellowship holders, provides funds for conference attendance, trains their communication skills, and supports them indeed whenever needed.

Thirty-seven percent (84 fellows) had left academic research either immediately after completing the Ph.D. or a couple of years later.[*] Since it is reasonable to assume that only scientists working in academia continuously publish their results (WEINGART, 2001, p. 91), our bibliometric analyses used only the publication lists of scientists that had worked without interruption in academic institutions.

All in all, 2,039 articles from 120 former fellowship recipients were included in our analyses.[**] The vast majority (98%) of the articles were published in English, and 2% were published in German or French. The articles were published in 508 different journals; in 36 journals, ten or more articles from fellows of the foundation had appeared (Table 7). According to the Institute for Scientific Information (ISI, Philadelphia, Pennsylvania, USA), in the year 2000 the impact factor of these journals (a measure of the frequency with which the "average article" in a journal has been cited in a particular year or period, revealing a journal's importance relative to others in its field) varied between 32.440 (*Cell*) and 2.461 (*Gene*).[***] By the end of 2001, the 2,039 articles published by the group of former fellowship recipients had been cited altogether 82,099 times.

How can we judge whether the citation rates for the publications by the Boehringer Ingelheim Fonds fellows are high or low? 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). For example, the "crown indicator" was computed as a measure of scientific impact in an international comparative bibliometric study on the scientific performance of German medical research carried out by CWTS on behalf of the German Federal Ministry of Education and Research (BMBF, Berlin, Germany) (TIJSSEN et al., 2002).

---

[*] ENDERS & BORNMANN (2001, p. 101) found similar results in a representative survey of the career paths of biologists that had completed doctorates at German universities in the mid to late 1980s (see also BORNMANN & ENDERS, 2004). One year after completing the Ph.D., 46% of the biologists that had been employed by a university or research institute during doctoral studies or had received a doctoral scholarship had left academic research.

[**] Of the 141 fellowship recipients with an up-to-date publication list, 21 could not be included in the analysis, as the available data was incomplete.

[***] In 2000, the highest impact factor in the ISI journal ranking list (considering all indexed journals) was achieved by the *Annual Review of Immunology* (50.340). On the list, *Cell* ranked third, *Nature* tenth and *Science* thirteenth.

Table 7. Journals in which ten or more articles by fellows of the Boehringer Ingelheim Fonds had been published after approval of their fellowships (ISI journal impact factor in 2000, $n = 2,009$. Thirty articles published in *B.I.F. Futura* are not included, as ISI does not index this journal)

Journal	Journal Impact Factor in 2000	Number of articles
Proceedings of the National Academy of Sciences USA	10.789	103
Journal of Biological Chemistry	7.360	95
EMBO Journal	13.999	74
Nature	28.689	60
Development (Cambridge, England)	9.353	59
Cell	32.440	41
Science	23.872	39
Journal of Cell Biology	13.955	36
FEBS Letters	3.440	34
Nucleic Acids Research	5.396	28
Current Biology	8.393	28
Journal of Virology	5.930	26
Journal of Neuroscience	8.502	25
Journal of Molecular Biology	5.388	24
European Journal of Immunology	5.240	22
Molecular and Cellular Biology	9.669	21
Mechanisms of Development	4.154	21
Journal of Immunology	6.834	20
Genes & Development	19.676	20
Journal of Experimental Medicine	15.236	19
Gene	2.461	19
European Journal of Biochemistry	2.852	19
Biochemistry	4.221	19
Journal of Neurochemistry	4.900	17
Journal of Cell Science	5.996	17
Genomics	3.425	14
Pharmacogenetics	4.465	13
Oncogene	6.490	13
Neuron	15.081	12
Infection and Immunity	4.204	12
Trends in Biochemical Sciences	13.246	10
Neuroreport	2.696	10
Human Molecular Genetics	9.048	10
European Journal of Neuroscience	3.862	10
Biochemical and Biophysical Research Communications	3.055	10
Other journals (altogether 472 different journals with less than ten articles each)	−	1,009

To determine the "crown indicator" for the publications by the Boehringer Ingelheim Fonds fellows, we divided the "mean number of citations for publications from fellowship recipients" by the "mean number of citations of all publications in the

journal sets chosen by the fellowship recipients". The quotient allows us to determine whether the citation impact of the fellowship recipient is far below (indicator value <0.5), below (indicator value 0.5–0.8), approximately the same as (0.8–1.2), above (1.2–1.5), or far above (>1.5) the international (primarily the Western world) citation impact baseline for the chosen journal sets. With ratio values above 1.5, the probability of identifying very good to excellent researchers is very high (Van Raan, 2004, pp. 31-32).
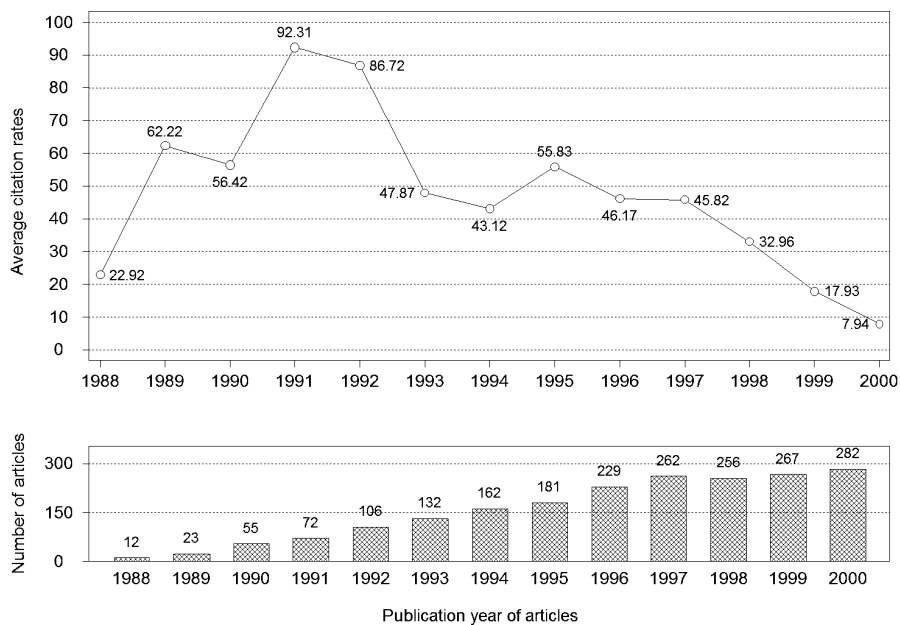

Figure 1. (Top) Mean number of citations of the articles published by the Boehringer Ingelheim Fonds fellows. (Bottom) Number of articles published in the year indicated.

Figure 1 (top) shows the mean number of citations of the articles published by the Boehringer Ingelheim Fonds fellows by the end of 2001. For example, each of the 72 articles published in 1991 was cited on average 92.31 times by the end of 2001, and each of the 282 articles published in 2000 was cited on average 7.94 times by the end of 2001. To calculate the "crown indicators", we used the ISI journal sets corresponding

to the fields "Multidisciplinary",* "Molecular Biology & Genetics"** and "Biology & Biochemistry".*** Out of the 22 ISI journal sets**** we chose "Molecular Biology & Genetics" and "Biology & Biochemistry" as reference sets, since 77% of the former fellowship recipients are biologists (61%) or biochemists (16%). Moreover, about a third of the research projects were in the field of molecular biology. In addition, we included the journal set "Multidisciplinary", since a large number of papers by Boehringer Ingelheim Fonds fellows were published in *The Proceedings of the National Academy of Sciences USA*, *Science* and *Nature* (Table 7),***** which are in this ISI category.

Table 8 lists the "crown indicators" of the publications classified according to journal set and year of publication. The values show that on average the papers by the fellowship recipients were significantly more frequently cited than the "average" publication in one of the three journals sets: 21 of the 30 "crown indicators", shown in Table 8, are above 1.5 (between 1.52 and 4.01), and seven are between 1.2 and 1.5.****** Only two values (0.96 and 1.02) are in the range that VAN RAAN (2004) denotes as "average". In the light of the mean citation rate achieved by the articles published by the Boehringer Ingelheim Fonds fellows, the decisions made by the foundation's Board have a high predictive validity.

---

* The "Multidisciplinary" category covers the spectrum of major scientific disciplines and includes journals of a broad or general character in the sciences (e.g., *Nature*, *Proceedings of the National Academy of Sciences USA*, *Science*) (INSTITUTE FOR SCIENTIFIC INFORMATION, 2002).

** "Molecular Biology & Genetics" journals include, for example, *Annual Review of Cell Biology*, *Cell*, *Annual Review of Cell and Developmental Biology* (INSTITUTE FOR SCIENTIFIC INFORMATION, 2002).

*** "Biology & Biochemistry" includes, for example, *Annual Review of Biochemistry*, *Physiological Reviews*, *Endocrine Reviews* (INSTITUTE FOR SCIENTIFIC INFORMATION, 2002).

**** Agricultural Sciences; Biology & Biochemistry; Chemistry; Clinical Medicine; Computer Science; Ecology/Environment; Economics & Business; Engineering; Geosciences; Immunology; Material Sciences; Mathematics; Microbiology; Molecular Biology & Genetics; Multidisciplinary; Neuroscience & Behavior; Pharmacology & Toxicology; Physics, Plant & Animal Science; Psychology/Psychiatry; Social Sciences, general; Space Science (INSTITUTE FOR SCIENTIFIC INFORMATION, 2002).

***** A comparison with other journal sets, for example "Clinical Medicine" or "Microbiology", shows that the "average" publication in the journal sets "Multidisciplinary", "Molecular Biology & Genetics" and "Biology & Biochemistry" has a much higher mean citation rate.

****** The average citation rates of articles published by the fellows between 1988 und 1990 are not listed in Table 8, since ISI no longer provides the corresponding average citation rates for papers published in those years.

Table 8. Average citation rates of papers published by recipients of Boehringer Ingelheim Fonds fellowships compared to mean citation rates of papers in the ISI journal sets "Multidisciplinary", "Molecular Biology & Genetics" and "Biology & Biochemistry" by publication year (1991–2000)

	Year of publication									
	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Journal set "Multidisciplinary"										
Mean number of citations for papers by fellowship recipients from the year of publication to 2001	92.31 (*n*=72)[*]	86.72 (*n*=106)	47.87 (*n*=132)	43.12 (*n*=162)	55.83 (*n*=181)	46.17 (*n*=229)	45.82 (*n*=262)	32.96 (*n*=256)	17.93 (*n*=267)	7.94 (*n*=282)
Baseline[**] for the journal set	50.00	47.31	49.86	42.17	44.40	36.71	31.09	21.67	14.75	6.33
Crown indicator (mean of citations divided by baseline)	1.85	1.83	0.96	1.02	1.26	1.26	1.47	1.52	1.22	1.25
Journal set "Molecular Biology & Genetics"										
Mean number of citations for papers by fellowship recipients from the year of publication to 2001	92.31 (*n*=72)[*]	86.72 (*n*=106)	47.87 (*n*=132)	43.12 (*n*=162)	55.83 (*n*=181)	46.17 (*n*=229)	45.82 (*n*=262)	32.96 (*n*=256)	17.93 (*n*=267)	7.94 (*n*=282)
Baseline[**] for the journal set	40.16	38.22	36.83	32.63	28.09	23.33	20.27	15.65	10.54	4.55
Crown indicator (mean of citations divided by baseline)	2.30	2.27	1.30	1.32	1.99	1.98	2.26	2.11	1.70	1.75
Journal set "Biology & Biochemistry"										
Mean number of citations for papers by fellowship recipients from the year of publication to 2001	92.31 (*n*=72)[*]	86.72 (*n*=106)	47.87 (*n*=132)	43.12 (*n*=162)	55.83 (*n*=181)	46.17 (*n*=229)	45.82 (*n*=262)	32.96 (*n*=256)	17.93 (*n*=267)	7.94 (*n*=282)
Baseline[**] for the journal set	23.04	22.30	20.79	19.24	16.44	13.89	11.85	8.88	5.77	2.56
Crown indicator (mean of citations divided by baseline)	4.01	3.89	2.30	2.24	3.40	3.32	3.87	3.71	3.11	3.10

[*] *n* = number of papers.
[**] Baselines are measures of cumulative citation frequencies across all papers published in a journal set: an average of 50.00 for the journal set "Multidisciplinary" in 1991 means that, on average, papers in "Multidisciplinary" journals were cited 50.00 times from 1991 to the end of 2001.

Conclusions

In this first comprehensive study on committee peer review for the selection of doctoral (Ph.D.) and post-doctoral research fellowship recipients, we analysed the selection procedure used by the Boehringer Ingelheim Fonds with regard to reliability, fairness and predictive validity.

In the analysis of reliability, the degree of agreement among reviewers was determined. In 76% of the cases, the decision on whether to award a fellowship or not was characterized by agreement. To characterise the extent of agreement or disagreement in the Board of Trustees of the Boehringer Ingelheim Fonds, we compared our results to the findings of other studies. It is important to take into consideration that in the other studies, the extent of agreement is not calculated indirectly by decision round, but directly by the level of agreement between two or more reviewers' ratings. For grant reviews, the following agreement coefficients are reported by other studies: in the selection procedure of the Deutsche Forschungs-gemeinschaft (Bonn, Germany), 82% of the reviewers' ratings are identical (HARTMANN & NEIDHARDT, 1990). According to CICCHETTI (1991), 68% of applications receive the same assessment in the peer review system of the National Science Foundation. HODGSON (1997) calculated an agreement rate of 73% for reviewers of the Heart and Stroke Foundation of Canada (Ottawa, Canada). Thus, the extent of agreement between reviewers, and thus the reliability of the committee peer review procedure of the Boehringer Ingelheim Fonds, is similar to that of major funding organizations.

With regard to fairness, we analysed whether potential sources of bias – gender, nationality, major field of study and institutional affiliation – could have influenced the fellowship award decisions. For post-doctoral fellowships, no statistically significant influence of any of these variables could be observed. For doctoral fellowships, we found evidence of an institutional, major field of study and gender bias, but not of a nationality bias. This incongruent result reflects the inconsistent findings of other empirical studies investigating the fairness of peer review. For example, some studies examining gender bias in review procedures indicate that women scientists are at a disadvantage (BROUNS, 2000; WENNERÅS & WOLD, 1997). However, a similar number of studies report only moderate effects or no gender effects (COLE, 1992; NATIONAL SCIENCE FOUNDATION, 2000; WARD & DONNELLY, 1998). An experimental study by SONNERT (1995, p. 47) found that grant submissions by women biologists received even better average evaluations than men's grant submissions did (mean rating: 3.67 vs. 3.27; $p = 0.0496$).

One principal problem that a survey of bias studies should take into account and that affects bias research in general is the lack of experimental studies. There have been only very few attempts to study reviewer bias directly in the natural setting of actual referee evaluations (ABRAMOWITZ et al., 1975; BAXT et al., 1998; MAHONEY, 1977; NYLENNA et al., 1994; PETERS & CECI, 1982). PETERS & CECI (1982), for instance, examined in a natural setting referees' evaluations of manuscripts submitted to American psychology journals (DUNCAN & MAGNUSON, 2003). They looked for reviewer bias that could be attributed to reviewers' knowledge of the authors' institutions or names. As test materials they selected already published research articles by investigators from prestigious and highly productive American psychology departments. With fictitious names and institutions substituted for the original ones, the altered manuscripts were formally resubmitted to the journals that had originally refereed and published them. Eight of the nine altered articles were rejected. Peters & Ceci's bias study was criticized, however, for having violated ethical principles for research with human subjects (CHUBIN, 1982; FLEISS, 1982; HONIG, 1982; WELLER, 2002). The lack of experimentally derived findings makes it impossible to establish unambiguously whether work from a particular group of scientists receives better reviews (and thus has a higher approval rate) due to biases in the review and decision-making procedure, or if favourable review and greater success in the selection procedure is simply a consequence of the scientific merit of the corresponding group of applicants.

The most important aspect of our study was to test the predictive validity of the review procedure, that is, whether the foundation achieves its goal to select the best junior scientists to receive fellowships. Our bibliometric analysis showed that this is indeed the case and that the selection procedure is thus highly valid: journal articles published by Boehringer Ingelheim Fonds fellows are cited considerably more often than the "average" publication in the ISI journal sets "Multidisciplinary", "Molecular Biology & Genetics", and "Biology & Biochemistry". These sets include journals covering the research fields in which most of the fellows publish. Similar results were reported for the decisions of the editors of the *Journal of Clinical Investigation* (WILSON, 1978), *British Medical Journal* (LOCK, 1985) and *Angewandte Chemie* (DANIEL, 1993): "Based on mean citation rates for accepted manuscripts and rejected manuscripts that were nevertheless published elsewhere, editorial decisions in all the existing studies reflect a high degree of predictive validity" (p. 56). In addition, CHAPMAN & MCCAULEY (1994) and MAVIS & KATZ (2003) reported similar findings for quality ratings of graduate fellows funded by the National Science

Foundation and for funding decisions of the March of Dimes Birth Defects Foundation (Indianapolis, IN, USA).[*]

All in all, the results show that the selection procedure implemented by the Boehringer Ingelheim Fonds is highly valid, meaning that it achieves its objective to select the best junior scientists to receive fellowships. However, our study found some evidence that three potential sources of bias (institutional affiliation, major field of study and gender) may influence the decisions of the Board of Trustees.

References

ABATE, T. (1995). What's the verdict on peer review? *Ethics in Research*, 1: 1.

ABRAMOWITZ, S. I., GOMES, B., ABRAMOWITZ, C. V. (1975). Publish or politic: referee bias in manuscript review. *Journal of Applied Social Psychology*, 3: 187–200.

BAXT, W. G., WAECKERLE, J. F., BERLIN, J. A., CALLAHAM, M. L. (1998). Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance. *Annals of Emergency Medicine*, 32: 310–317.

BLOSSFELD, H.-P., ROHWER, G. (2002). *Techniques of Event History Modeling. New Approaches to Causal Analysis.* Mahwah, NJ, USA: Lawrence Erlbaum Associates.

BORNMANN, L. (2004). *Stiftungspropheten in der Wissenschaft. Zuverlässigkeit, Fairness und Erfolg des Peer-Review.* Münster: Waxmann.

BORNMANN, L., DANIEL, H.-D. (2003). Begutachtung durch Fachkollegen in der Wissenschaft. Stand der Forschung zur Reliabilität, Fairness und Validität des Peer-Review-Verfahrens. In: SCHWARZ, S., TEICHLER, U. (Eds), *Universität auf dem Prüfstand. Konzepte und Befunde der Hochschulforschung.* Frankfurt am Main: Campus, pp. 211–230.

BORNMANN, L., 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. *B.I.F. Futura*, 19: 7–19.

BORNMANN, L., ENDERS, J. (2004). Social origin and gender of doctoral degree holders. Impact of particularistic attributes in access to and in later career attainment after achieving the doctoral degree in Germany. *Scientometrics*, 61: 19–41.

BROUNS, M. (2000). The gendered nature of assessment procedures in scientific research funding: the Dutch case. *Higher Education in Europe*, 25: 193–199.

CAMPANARIO, J. M. (1998a). Peer review for journals as it stands today – part 1. *Science Communication*, 19: 181–211.

---

[*] Although according to SHADISH (1989) "of all the science indicators we have, only citation counts are widely available, inexpensive, intuitively plausible, perceived to be reasonably fair, and generally applicable to the scientific community and its products" (p. 394), we plan to consider further success rate factors in addition to bibliometric indicators in determining the predictive validity of the B.I.F. peer review procedure. For example, the administrative office of the foundation has some information on the further career paths of the fellows. However, for conducting retrospective event history analysis (BLOSSFELD & ROHWER, 2002; ENDERS & BORNMANN, 2001) the B.I.F. database lacks detailed information on the various stages of the fellows' careers (such as type of employment, start and end dates for individual periods of employment, sector of employment). As the evaluation of career course data would provide a good complement to the bibliometric analyses, we plan in a future study to conduct a survey of the fellows in order to gather the needed data on their career paths. Recently, the WELLCOME TRUST (2001, London, UK), for example, conducted an in-depth study that followed the career paths of Trust-funded individuals.

CAMPANARIO, J. M. (1998b). Peer review for journals as it stands today – part 2. *Science Communication*, 19: 277–306.

CHAPMAN, G. B., MCCAULEY, C. (1994). Predictive validity of quality ratings of National Science Foundation graduate fellows. *Educational and Psychological Measurement*, 54: 428–438.

CHUBIN, D. E. (1982). Reforming peer-review – from recycling to reflexivity. *Behavioral and Brain Sciences*, 5: 204.

CICCHETTI, D. V. (1991). The reliability of peer-review for manuscript and grant submissions – a cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14: 119–134.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74: 829–836.

COLE, J. R. (2000). A short history of the use of citations as a measure of the impact of scientific and scholarly work. In: CRONIN, B., ATKINS, H. B. (Eds), *The Web of Knowledge. A Festschrift in Honor of Eugene Garfield*. Medford, New Jersey, USA: Information Today, pp. 281–300.

COLE, S. (1992). *Making Science. Between Nature and Society*. Cambridge, MA, USA: Harvard University Press.

COLE, S., COLE, J. R. (1967). Scientific output and recognition – study in operation of reward system in science. *American Sociological Review*, 32: 377–390.

COLE, S., FIORENTINE, R. (1991). Discrimination against women in science: the confusion of outcome with process. In: ZUCKERMAN, H., COLE, J. R., BRUER, J. T. (Eds), *The Outer Circle. Women in the Scientific Community*. London, UK: W W Norton & Company, pp. 205–226.

CONROY, R. M. (2002). Choosing an appropriate real-life measure of effect size: the case of a continous predictor and a binary outcome. *The Stata Journal*, 2: 290–295.

DANIEL, H.-D. (1993). *Guardians of Science. Fairness and Reliability of Peer Review*. Chichester, UK: John Wiley & Sons, Ltd.

DEMICHELI, V., PIETRANTONJ, C. (2004). Peer review for improving the quality of grant applications (Cochrane Methodology Review). In: *The Cochrane Library, Issue 1*. Chichester, UK: John Wiley & Sons, Ltd.

DUNCAN, G. J., MAGNUSON, K. A. (2003). The promise of random-assignment social experiments for understanding well-being and behavior. *Current Sociology*, 51: 529–541.

EISENHART, M. (2002). The paradox of peer review: admitting too much or allowing too little? *Research in Science Education*, 32: 241–255.

ENDERS, J., BORNMANN, L. (2001). *Karriere mit Doktortitel? Ausbildung, Berufsverlauf und Berufserfolg von Promovierten*. Frankfurt am Main: Campus.

FINN, C. E. (2002). The limits of peer review. *Education Week*, 21: 30–34.

FLEISS, J. L. (1982). Deception in the study of the peer-review process. *Behavioral and Brain Sciences*, 5: 210–211.

FRÖHLICH, H. (2001). It all depends on the individuals. Research promotion – a balanced system of control. *B.I.F. Futura*, 16: 69–77.

GODLEE, F., DICKERSIN, K. (2003). Bias, subjectivity, chance, and conflict of interest. In: GODLEE, F., JEFFERSON, J. (Eds), *Peer review in health sciences*. London: BMJ Publishing Group, pp. 91–117.

GOODMAN, S. N., BERLIN, J., FLETCHER, S. W., FLETCHER, R. H. (1994). Manuscript quality before and after peer-review and editing at Annals of Internal-Medicine. *Annals of Internal Medicine*, 121: 11–21.

HACKETT, E. J., CHUBIN, D. E. (2003). Peer review for the 21st century: applications to education research. Paper presented at the conference entitled *Peer Review of Education Research Grant Applications. Implications, Considerations, and Future Directions*, Washington, DC, USA.

HAGSTROM, W. O. (1965). *The Scientific Community*. New York, NY, USA: Basic Books.

HARTMANN, I., NEIDHARDT, F. (1990). Peer-review at the Deutsche Forschungsgemeinschaft. *Scientometrics*, 19: 419–425.

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: 1189–1195.

Hojat, M., Gonnella, J. S., 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: 75–96.

Honig, W. M. (1982). Peer-review in the physical sciences – an editors view. *Behavioral and Brain Sciences*, 5: 216–217.

Horrobin, D. F. (2001). Something rotten at the core of science? *Trends in Pharmacological Sciences*, 22: 51–52.

Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression.* Chichester, UK: John Wiley & Sons, Inc.

Ince, M. (1991). US research may drop peer review for lottery. *Times Higher Education Supplement*, 955: 85.

Institute for Scientific Information (2002). *ISI Essential Science Indicators v1.0.* Philadelphia, PA, USA: Institute for Scientific Information (ISI).

Jayasinghe, U. W., Marsh, H. W., Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society Series A – Statistics in Society*, 166: 279–300.

Klahr, D. (1985). Insiders, outsiders, and efficiency in a National Science Foundation panel. *American Psychologist*, 40: 148–154.

Kostoff, R. N. (1997). The principles and practices of peer review. *Science and Engineering Ethics*, 3: 19–34.

Langfeldt, L. (2004). Expert panels evaluating research: decision-making and sources of bias. *Research Evaluation*, 13: 51–62.

Lock, S. (1985). *A difficult Balance: Editorial Peer Review in Medicine.* Philadelphia, PA, USA: ISI Press.

Mahoney, M. J. (1977). Publication prejudices: an experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, 2: 161–175.

Mander, A., Clayton, D. (1999). Hotdeck imputation. *Stata Technical Bulletin*, 51: 16–18.

Mavis, B., Katz, M. (2003). Evaluation of a program supporting scholarly productivity for new investigators. *Academic Medicine*, 78: 757–765.

McClellan, J. E. (2003). Specialist control – The publications Committee of the Academie-Royal-des-Sciences (Paris) 1700–1793. *Transactions of the American Philosophical Society*, 93: VII.

Merton, R. K. (1948). The self-fulfilling prophecy. *Antioch Review*, 8: 193–210.

Merton, R. K. (1968). Matthew effect in science. *Science*, 159: 56–63.

Moran, G. (1998). *Silencing Scientists and Scholars in Other Fields: Power, Paradigm Controls, Peer Review, and Scholarly Communication.* London, UK: Ablex Publishing Corporation.

Mosteller, F., Tukey, J. W. (1977). *Data Analysis and Regression. A Second Course in Statistics.* Boston, MA, USA: Addison-Wesley.

National Science Foundation (2000). *Report to the National Science Board on the National Science Foundation's Merit Review System Fiscal Year 1999.* Arlington, VA, USA: National Science Foundation.

Nylenna, M., Riis, P., Karlsson, Y. (1994). Multiple blinded reviews of the 2 manuscripts – effects of referee characteristics and publication language. *Journal of the American Medical Association*, 272: 149–151.

Overbeke, J., Wager, E. (2003). The state of the evidence: what we know and what we don't know about journal peer review. In: Godlee, F., Jefferson, T. (Eds), *Peer Review in Health Sciences*. London, UK: BMJ Books, pp. 45–61.

Owen, R. (1982). Reader bias. *Journal of the American Medical Association*, 247: 2533–2534.

Peters, D. P., Ceci, S. J. (1982). Peer-review practices of psychological journals – the fate of accepted, published articles, submitted again. *Behavioral and Brain Sciences*, 5: 187–195.

Pierie, J. P. E. N., Walvoort, H. C., 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: 1480–1483.

Polanyi, M. (1966). *The Tacit Dimension.* New York, NY, USA: Doubleday.

Pʀᴜᴛʜɪ, S., Jᴀɪɴ, A., Wᴀʜɪᴅ, A., Mᴇʜʀᴀ, K., Nᴀʙɪ, S. A. (1997). Scientific community and peer review system – a case study of a central government funding scheme in India. *Journal of Scientific & Industrial Research*, 56: 398–407.

Rᴀʙᴇ-Hᴇsᴋᴇᴛʜ, S., Eᴠᴇʀɪᴛᴛ, B. (2004). *A Handbook of Statistical Analyses Using Stata.* Boca Raton, UK: Chapman & Hall/CRC.

Rᴇɴɴɪᴇ, D. (2003). Innovation and peer review. In: Gᴏᴅʟᴇᴇ, F., Jᴇꜰꜰᴇʀsᴏɴ, T. (Eds), *Peer Review in Health Sciences.* London, UK: BMJ Books, pp. 76–90.

Rᴏss, P. F. (1980). *The Sciences' Self-Management: Manuscript Refereeing, Peer Review, and Goals in Science.* Massachusetts, MA, USA: The Ross Company, Todd Pond.

Rᴏʏ, R. (1985). Funding science – the real defects of peer-review and an alternative to it. *Science Technology & Human Values*, 52: 73–81.

Rᴜʙɪɴ, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Chichester, UK: John Wiley & Sons, Ltd.

Rᴜʙɪɴ, D. B., Sᴄʜᴇɴᴋᴇʀ, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81: 366–374.

Sᴄʜᴀꜰᴇʀ, J. L. (2000). *Analysis of Incomplete Multivariate Data by Simulation.* London, UK: Chapman and Hall.

Sʜᴀᴅɪsʜ, W. R. (1989). The perception and evaluation of quality in science. In: Gʜᴏʟsᴏɴ, B., Sʜᴀᴅɪsʜ, W. R., Nᴇɪᴍᴇʏᴇʀ, R. A., Hᴏᴜᴛs, A. C. (Eds), *Psychology of Science. Contributions to Metascience.* Cambridge, UK: Cambridge University Press, pp. 383–426.

Sʜᴀʀᴘ, D. W. (1990). What can and should be done to reduce publication bias – the perspective of an editor. *Journal of the American Medical Association*, 263: 1390–1391.

Sᴏɴɴᴇʀᴛ, G. (1995). What makes a good scientist? Determinants of peer evaluation among biologists. *Social Studies of Science*, 25: 35–55.

SᴛᴀᴛᴀCᴏʀᴘ. (2003). *Stata Statistical Software: Release 8.* College Station, Texas, USA: Stata Corporation.

Tɪᴊssᴇɴ, R. J. W., ᴠᴀɴ Lᴇᴇᴜᴡᴇɴ, T. N., ᴠᴀɴ Rᴀᴀɴ, A. F. J. (2002). *Mapping the Scientific Performance of German Medical Research. An International Comparative Bibliometric Study.* Stuttgart: Schattauer.

Uɴɪᴛᴇᴅ Sᴛᴀᴛᴇs Gᴇɴᴇʀᴀʟ Aᴄᴄᴏᴜɴᴛɪɴɢ Oꜰꜰɪᴄᴇ (1999). *Peer Review Practices at Federal Science Agencies Vary.* Washington, DC, USA: United States General Accounting Office.

ᴠᴀɴ Rᴀᴀɴ, A. F. J. (1999). Advanced bibliometric methods for the evaluation of universities. *Scientometrics*, 45: 417–423.

Vᴀɴ Rᴀᴀɴ, A. F. J. (2004). Measuring science. Capita selecta of current main issues. In: Mᴏᴇᴅ, H. F., Gʟäɴᴢᴇʟ, W., Sᴄʜᴍᴏᴄʜ, U. (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems.* Dordrecht: Kluwer Academic Publishers, pp. 19–50.

Wᴀʀᴅ, J. E., Dᴏɴɴᴇʟʟʏ, N. (1998). Is there gender bias in research fellowships awarded by the NHMRC? *Medical Journal of Australia*, 169: 623–624.

Wᴇɪɴɢᴀʀᴛ, P. (2001). *Die Stunde der Wahrheit? Zum Verhältnis der Wissenschaft zu Politik, Wirtschaft und Medien in der Wissensgesellschaft.* Weilerswist: Velbrück.

Wᴇʟʟᴄᴏᴍᴇ Tʀᴜsᴛ (2001). *Review of Wellcome Trust PhD Research Training. Career Paths of a 1988–1990 Prize Student Cohort.* London, UK: Wellcome Trust.

Wᴇʟʟᴇʀ, A. C. (2002). *Editorial Peer Review: Its Strengths and Weaknesses.* Medford, New Jersey, USA: Information Today, Inc.

Wᴇɴɴᴇʀås, C., Wᴏʟᴅ, A. (1997). Nepotism and sexism in peer-review. *Nature*, 387: 341–343.

Wᴇssᴇʟʏ, S. (1998). Peer review of grant applications: what do we know? *Lancet*, 352: 301–305.

Wɪʟsᴏɴ, J. D. (1978). Peer review and publication. *Journal of Clinical Investigation*, 61: 1697–1701.

Wᴏᴏᴅ, F. Q., Wᴇssᴇʟʏ, S. (2003). Peer review of grant applications: a systematic review. In: Gᴏᴅʟᴇᴇ, F., Jᴇꜰꜰᴇʀsᴏɴ, T. (Eds), *Peer Review in Health Sciences.* London, UK: BMJ Books, pp. 14–44.